Philippe Chartier
Even Loarer

Évaluer l'intelligence logique

Approche cognitive et dynamique

ÉCHELLES D'INTELLIGENCE
(WISC-III, WISC-IV, WAIS III)

TESTS DE FACTEUR G
(Raven, dominos…)

BATTERIES FACTORIELLES
(NV5, NV7, DAT5)

DUNOD

Table des matières

Introduction

LES tests d'intelligence datent du début du XX^{e}. Depuis cette époque de très nombreux ouvrages sont parus sur le sujet, à destination des chercheurs, des étudiants et/ou des praticiens. Alors... pourquoi un livre de plus ?

La première ambition de cet ouvrage est d'apporter une **vision à la fois large et actualisée** de l'évaluation de l'intelligence logique, qui intègre à la fois la présentation des « standards » classiques et celle des évolutions plus récentes dans le domaine, et cela, tant du point de vue des connaissances théoriques que des méthodes et outils d'évaluation.

La seconde ambition est qu'il fournisse une aide et un **soutien théorique et méthodologique au travail du praticien** dans toutes les étapes et dans tous les aspects du processus d'évaluation de l'intelligence logique :

- pour le choix des épreuves (selon les objectifs, les personnes, les contraintes et conditions de passation, la qualité des épreuves et des étalonnages...) ;
- pour la mise en œuvre de l'évaluation (en temps libre ou limité, en individuel ou collectif...) ;
- pour la correction et l'interprétation des résultats (indices, étalonnages, scatters, mise en relation avec des critères...) ;
- pour la restitution aux personnes évaluées (manière de le faire, supports...) ;
- pour le respect des règles de déontologie et la mise en ouvre de pratiques non discriminatoires.

La troisième ambition est qu'il puisse être un bon **support pédagogique à la formation** des étudiants de psychologie dans ce domaine. La place accordée à l'enseignement de la méthodologie de la mesure en psychologie, et en particulier à la formation à la méthode des tests, est assez hétérogène selon les universités, alors même que l'on observe depuis quelques années un fort regain d'intérêt des praticiens, et futurs praticiens, pour ce domaine,

mais aussi une augmentation de la demande sociale et de celle des particuliers, en manière d'évaluation.

Une formation solide à la pratique des tests est d'autant plus importante dans le contexte actuel où les pratiques d'évaluation, en particulier via Internet, mais également dans de nombreux cabinets spécialisés, ne satisfont souvent pas aux critères qui leur garantissent un minimum de validité[1]. Les pratiques peu valides ne préservent pas le droit des personnes évaluées à bénéficier d'un traitement équitable, chaque fois qu'une décision est prise sur la base de ces évaluations. Elles trompent également la personne qui cherche plus simplement à « mieux se connaître ».

L'ouvrage vise donc à faire le point sur les principaux éléments théoriques et méthodologiques sur lesquels reposent les pratiques d'évaluation de l'intelligence logique. Il dresse un panorama des tests dans ce domaine et fournit un certain nombre d'indications concernant les spécificités, qualités, utilisations et limites de ces différents tests[2]. Plus précisément, nous avons souhaité présenter :

1. Les cadres historiques, théoriques, méthodologiques et déontologiques qui nous semblent indispensables pour garantir la fiabilité d'une évaluation psychologique ;
2. Une large sélection d'épreuves utilisables en France, certaines déjà largement connues et utilisées (échelles de Wechsler, Matrices de Raven...), d'autres sans doute moins (le logiciel *Samuel,* les pratiques d'évaluation dynamique...), afin de regrouper, dans un même ouvrage, un ensemble assez vaste d'outils aujourd'hui disponibles et utilisables. Dans la mesure du possible, nous avons illustré ces épreuves par des exemples d'items[3] ;
3. Une analyse de ces épreuves. Il ne s'agissait pas pour nous de lister uniquement des épreuves mais d'apporter, en toute modestie et en nous appuyant sur leur analyse et sur l'expérience de leur mise en œuvre, un regard critique et des suggestions et recommandations sur ces outils et

1. En la matière, le pire côtoie souvent le meilleur et le peu de communication des sociétés sur les méthodes utilisées, sous couvert de protection concurrentielle, ne permet souvent pas de faire un choix éclairé.

2. Ce qui le distingue par exemple de l'ouvrage de Zurfluh (1976) qui était certes exhaustif, mais ne fournissait qu'une information limitée sur chaque test. En outre, cet ouvrage ne constitue plus aujourd'hui, du fait de son ancienneté, une référence suffisante à la pratique.

3. Nous remercions vivement les ECPA pour leur aimable autorisation de reproduire certains exemples d'items de tests

leurs usages. Dans tous les cas, le présent ouvrage n'est pas destiné à remplacer les manuels d'utilisation de ces épreuves. Nous souhaitons, au contraire qu'il renforce l'envie de s'y reporter et qu'il constitue également une invitation à la consultation de documents complémentaires (livres, articles, ...) relatifs aux approches et épreuves que nous présentons ;

4. D'autres approches évaluatives relativement méconnues, telles que l'analyse des stratégies de résolution ou encore l'évaluation dynamique de l'intelligence, qui apportent des perspectives de renouvellement des pratiques (et des outils) d'évaluation (Huteau et Lautrey, 1999a). Même si ces épreuves sont encore rares, et qu'elles restent souvent perfectibles, elles témoignent de rapprochements intéressants entre théories et pratiques et peuvent apporter des solutions pratiques très utiles à certaines problématiques.

Cet ouvrage présente bien entendu certaines limites. Il est limité dans son périmètre : centré sur la question de l'évaluation de l'intelligence logique il n'aborde pas la question de l'évaluation d'autres formes d'intelligence (sociale, pratique, émotionnelle...). Il est également limité dans les niveaux d'âges pris en compte : il concerne l'évaluation des adolescents et adultes et ne présente donc pas les épreuves utilisables auprès des enfants d'âge préscolaire et scolaire. Enfin, tous les test d'intelligence logique n'y figurent pas, par nécessité de faire des choix (par exemple les tests sur support verbal, tels que ceux élaborés par Bonnardel (*cf.* Thiébaut, 2000, pour une présentation), ou encore les tests inspirés de la théorie de Piaget (mieux adaptés pour les plus jeunes).

L'ouvrage est organisé en trois parties :

1. La première partie présente les **aspects historiques et théoriques** de l'intelligence logique (chapitre 1) ainsi que les **principes méthodologiques de sa mesure** (chapitre 2) ;
2. La seconde partie est consacrée à la **présentation des principales familles de tests d'intelligence** : les échelles d'intelligence (chapitre 3), les tests de facteur g (chapitre 4) et les batteries factorielles (chapitre 5) ;
3. La troisième partie porte sur l'**utilisation des tests**. Au-delà des approches classiques d'utilisation des tests qui consistent à recueillir et analyser des scores de performance et qui sont largement évoquées lors de la présentation des épreuves, deux orientations plus contemporaines de l'évaluation sont présentées dans cette partie : l'analyse des stratégies de résolution (chapitre 6) et l'approche de l'évaluation dynamique

(chapitre 7). Enfin un dernier chapitre est consacré aux différents cadres d'utilisation des tests ainsi qu'aux aspects déontologiques relatifs à l'évaluation des personnes.

En annexe figurent des fiches synthétiques des épreuves présentées ainsi que la reproduction de deux documents : le *Code de Déontologie des Psychologues* et les *recommandations internationales sur l'utilisation des tests*.

Conçu comme un manuel pratique, chacune de ses parties peut être lue de façon indépendante. De nombreux renvois sont faits dans le texte pour permettre de multiples itinéraires de lecture.

La pratique de l'évaluation nécessite à nos yeux la maîtrise d'un ensemble de connaissances souples, et articulées, ainsi qu'une pratique réflexive. Il est essentiel de ne pas appliquer de façon mécanique des procédures standard (comme par exemple le calcul des scores) mais de comprendre et maîtriser tous les aspects de l'évaluation (des conditions de standardisation à l'interprétation des scores) afin de pouvoir s'ajuster au mieux à chaque situation prise dans sa complexité, sans pour autant mettre en péril les principes de la standardisation. Cela est nécessaire pour s'assurer à la fois de la validité de la mesure et des conditions de respect des droits de la personne évaluée.

Si cet ouvrage contribue à l'un des objectifs fixés dans les *recommandations internationales sur l'utilisation des tests* de « promouvoir une bonne utilisation des tests et d'encourager des pratiques exemplaires dans le domaine de l'évaluation » (SFP, 2003, p. 9), nous pourrons considérer avoir atteint notre but.

PREMIÈRE PARTIE

Aspects historiques, théoriques et méthodologiques

CHAPITRE
1

Les conceptions théoriques de l'intelligence et de sa mesure

Sommaire

1. Définir et mesurer l'intelligence

Comprendre la nature et les propriétés de l'intelligence humaine est l'une des grandes préoccupations de la psychologie depuis ses origines.

La notion a fait l'objet de nombreuses tentatives de modélisation et a été au centre de nombreux débats, tant théoriques ou méthodologiques qu'idéologiques.

Elle a également donné lieu à la production d'un grand nombre de méthodes et d'outils d'évaluation.

Dans ce premier chapitre, nous présenterons les principaux modèles de l'intelligence proposés par différents auteurs tout au long du XX^{e} siècle. Nous verrons qu'ils ont été conçus dans certains contextes scientifiques, mais aussi sociaux et idéologiques et sont associés, dans la majorité des cas à des techniques et modalités d'évaluation dont les principales seront présentées dans la suite de cet ouvrage.

Définir l'intelligence

La diversité des modèles produits et des approches retenues par les auteurs témoigne de la difficulté rencontrée à cerner cette notion. L'intelligence humaine est en effet une abstraction. C'est un construit théorique élaboré pour rendre compte d'un ensemble de conduites humaines perçues comme efficientes.

Etymologiquement, le terme vient du latin *intelligere* qui signifie *comprendre*. Mais la simple fonction de comprendre ne suffit pas à l'évidence à rendre compte de l'intelligence humaine.

En 1921, soucieux d'y voir plus clair, les éditeurs du *Journal of Educational Psychology* demandèrent à un groupe d'experts reconnus dans le domaine de la psychologie de donner une définition de l'intelligence. Il en résulta une grande variété de réponses.

Réitérant l'exercice soixante-cinq ans plus tard, Sternberg et Detterman (1986) firent le même constat d'une absence de consensus. Néanmoins, dans ces deux études, ainsi que dans une troisième (menée un an plus tard par deux chercheurs américains (Snyderman et Rothman, 1987) selon un principe analogue auprès d'un large échantillon de plusieurs centaines d'experts), les caractéristiques présentant le plus fort consensus concernent les capacités

à mener des raisonnements abstraits, à résoudre des problèmes nouveaux, à acquérir de nouvelles connaissances, à s'adapter à l'environnement (*cf.* tableau 1.1). Viennent ensuite les capacités de mémorisation, de vitesse mentale, les capacités linguistiques et mathématiques ou encore la culture générale et la créativité. L'intelligence serait donc principalement comprise comme ce qui permet de comprendre, connaître, raisonner et résoudre des problèmes.

Tableau 1.1
Classement des caractéristiques essentielles de l'intelligence selon un échantillon de 661 experts (Étude de Snyderman et Rothman 1987).

1.	Pensée ou raisonnement abstrait	99 %
2.	Aptitude à résoudre des problèmes	98 %
3.	Capacité à acquérir des connaissances	96 %
4.	Mémoire	81 %
5.	Adaptation à l'environnement	77 %
6.	Vitesse mentale	72 %
7.	Capacité linguistique	71 %
8.	Capacité en mathématiques	68 %
9.	Culture générale	62 %
10.	Créativité	60 %

On peut cependant constater, plus généralement, que la finalité adaptative de l'intelligence est présente dans la grande majorité des définitions, comme l'indiquait déjà Wechsler en 1944 « l'intelligence est la capacité complexe ou globale d'un individu d'agir en fonction d'un but, de penser rationnellement et d'avoir des rapports efficaces avec son environnement » (cité par Grégoire, 2004, p. 150) ou encore Piaget en 1970 en affirmant que : « l'intelligence c'est l'adaptation ». Cette définition, trop générale pour pouvoir être fausse, ne doit cependant pas masquer les divergences entre auteurs que nous avons évoquées.

Plusieurs explications peuvent être données à ces divergences (*cf.* Lautrey, 2006). Une première serait de considérer que l'intelligence est une notion trop générale, trop floue, trop abstraite (comme il en est de même actuellement, par exemple, pour la notion de compétence), pour donner lieu à une véritable investigation scientifique, seule voie possible à l'élaboration

d'un consensus. La seconde, qui a la préférence de nombreux chercheurs contemporains en psychologie cognitive et différentielle, est que l'intelligence est une fonction adaptative de haut niveau qui se manifeste à travers une multiplicité de mécanismes et qui peut donc être appréhendée sous de très nombreux angles. Cianciolo et Sternberg (2004) illustrent cette position par la célèbre fable bouddhiste des aveugles et de l'éléphant : chacun en touche une partie différente et conclut que l'éléphant a les propriétés de la partie qu'il découvre. L'unité de l'intelligence peut-elle émerger de la somme des modèles qui la décrivent ? Ce n'est probablement pas si simple, car la question principale est celle de l'intégration des différentes fonctions et processus. Néanmoins, des travaux existent qui visent à proposer des visions synthétiques de plusieurs modèles[1].

Mesurer l'intelligence

La mesure de l'intelligence a, depuis ses premières origines, servi deux objectifs distincts bien qu'étroitement complémentaires.

– *Le premier objectif est d'ordre épistémologique*. Il concerne la production de connaissances sur ce qu'est l'intelligence humaine. Dans ce domaine comme dans beaucoup d'autres, la construction d'une connaissance scientifique et la mesure des phénomènes concernés sont, comme l'a souligné Bachelard (1934, 1938), étroitement liées. Évoquant le développement des sciences, Ullmo (1969, cité par Gillet, 1987) rappelle qu'« un pas décisif a été franchi lorsqu'on a compris que c'est la mesure qui définit la grandeur à mesurer, celle-ci ne préexiste pas à sa mesure, comme une intuition sommaire l'a fait longtemps croire » (p. 24).

– *Le second objectif est d'ordre pratique*. Il correspond à un besoin d'apporter des réponses à des demandes sociales. Dans l'histoire contemporaine des recherches sur l'intelligence, c'est souvent le second objectif qui a précédé le premier. C'est par exemple le cas des travaux de Binet. Si Binet est légitimement considéré comme l'un des pères de l'intelligence, il est remarquable de noter que son souci premier n'était pas de définir et modéliser l'intelligence mais de trouver des solutions pour mieux scolariser

1. C'est par exemple le cas des travaux au sein de l'approche factorielle qui proposent des modèles hiérarchiques synthétiques (Gustaffson (1984), Caroll (1993), ou encore de ceux de Lautrey (2001) qui rapprochent le courant psychométrique classique et l'étude des processus cognitifs. On peut également mentionner la tentative récente de Rozencwajg (2005) de proposer une vision intégrative de l'intelligence.

les enfants présentant des difficultés et des retards d'apprentissage et des déficits de réussite scolaire (Binet, 1911).

Ainsi, la mesure, forme standardisée et instrumentée d'observation, est nécessaire pour construire la connaissance. Mais la connaissance est également nécessaire à l'élaboration de méthodes et d'outils de mesure. Le paradoxe est bien là : construire de bons instruments de mesure nécessite une bonne connaissance de ce que l'on cherche à mesurer mais cette connaissance est elle-même dépendante des méthodes et instruments de mesure disponibles. Ce n'est donc que par un ajustement progressif et souvent laborieux de ces deux approches que la connaissance progresse.

En outre, l'élaboration théorique et la construction d'instruments de mesure ne se font qu'en fonction d'un certain contexte intellectuel et social. Celui-ci évoluant dans le temps, les définitions et conceptions de l'intelligence ont également évolué. Nous donnerons dans ce chapitre un rapide aperçu des principales étapes de cette évolution et des conceptions proposées par différents auteurs et évoquerons, lorsqu'elles existent les méthodes de mesure correspondantes.

2. Repères historiques

Les premiers tests mentaux

Les premières tentatives de mesure quantitative des processus mentaux sont apparues à la fin du XIX^{e} siècle avec la naissance de la psychologie scientifique.

Dans cette perspective, Wilhem Wund (1932-1920), psychologue allemand, crée à Leipzig en 1879 le premier Laboratoire de psychologie expérimentale. Il développe des méthodes précises de mesure des seuils perceptifs et des temps de réactions et cherche à comprendre les processus à l'œuvre dans ces tâches sensorielles élémentaires De nombreux étudiants européens et nord-américains viendront se former dans son laboratoire aux méthodes de la psychologie expérimentale. L'un des étudiants, venu des États-Unis est James McKeen Cattel (1960-1944). Alors que Wund est essentiellement préoccupé par l'établissement de lois générales des processus sensoriels, Cattel s'intéresse aux différences entre les individus et constate que celles-ci ont tendance à présenter une certaine stabilité. De retour aux États-Unis, il sélectionne certaines situations expérimentales et les utilise

pour étudier les différences interindividuelles. En 1890 il utilise le terme de « test mental » pour désigner ces situations expérimentales standardisées.

À la même époque, en Angleterre, Francis Galton (1822-1911), qui est cousin de Darwin, fait également des travaux sur les différences interindividuelles dans les processus sensoriels élémentaires. Il le fait dans l'optique de vérifier que la théorie de l'évolution de Darwin s'applique aussi au développement de l'intelligence dans l'espèce humaine. Galton crée des tests physiques et sensoriels qu'il applique de façon standardisée à de grands échantillons et invente les étalonnages. Il étudie les performances des parents et des enfants dans l'optique de montrer que les différences individuelles sont héréditaires et élabore, à cette occasion, les principes de la régression et du coefficient de corrélation.

Si les premiers tests mentaux ont été créés en fonction de préoccupations essentiellement scientifiques (comprendre les lois de la perception, tester la théorie de Darwin...), il est rapidement apparu qu'ils étaient susceptibles de contribuer à répondre à certains besoins de la société de l'époque.

La fin du XIX^{e} siècle est marquée par une forte industrialisation et par une volonté de généraliser l'éducation. De nouveaux besoins en découlent en matière d'évaluation des personnes à des fins d'orientation vers des formations ou vers des emplois. En France, les lois Ferry de 1881 et 1882 rendant l'instruction élémentaire obligatoire, ont fait émerger d'autres besoins d'évaluation, en particulier celui de distinguer parmi les élèves d'école élémentaire, ceux qui n'avaient pas les moyens intellectuels pour suivre l'enseignement général et leur fournir un enseignement adapté afin de remédier à ces retards de développement.

C'est en réponse à cette demande sociale qu'Alfred Binet a été amené à créer son test d'intelligence avec Théodore Simon.

De la mesure des processus élémentaires à celle des fonctions supérieures

Binet était très critique vis-à-vis des tests issus d'expériences de laboratoire et portant sur des processus élémentaires comme moyen d'évaluer les capacités intellectuelles qu'il percevait comme plus complexes. D'ailleurs, les premières tentatives de Cattel, d'évaluer à l'aide de ses tests mentaux les étudiants de l'université de Columbia donnèrent raison à Binet : elles démontrèrent qu'il n'y avait pas de relation entre les résultats dans ces tests et la réussite dans les études universitaires.

Binet était convaincu que les différences individuelles dans les capacités intellectuelles seraient mieux estimées par des tâches de mémorisation, de raisonnement, de jugement ou d'imagerie mentale. Encore fallait-il concevoir les tâches adaptées.

Binet commence alors avec Simon, qui était médecin dans un institut pour enfants « retardés », à mettre au point des épreuves nouvelles et à les essayer dans les écoles. Ils constatent que certains items échoués par des enfants « retardés » sont réussis par des enfants « normaux » de même âge. La réussite à ces items, ou groupes d'items, doit alors permettre de diagnostiquer un retard, ou une avance, de développement intellectuel. Chaque enfant peut ainsi être caractérisé à la fois par son âge réel et par un âge mental correspondant à son niveau de réussite (voir dans le chapitre 3 la présentation de la notion d'âge mental).

En 1904, une commission ministérielle, la commission Bourgeois, charge officiellement Binet d'étudier le problème du diagnostic de la débilité mentale. Il ne mettra, avec Simon qu'un an à mettre au point leur première échelle métrique de d'intelligence. Nous reviendrons plus en détail sur cette échelle dans le chapitre 3.

L'échelle de Binet-Simon a eu un succès immédiat et fulgurant. Elle permettait de sortir de l'impasse où se trouvait le problème de l'évaluation de l'intelligence et fournissait enfin des moyens de répondre aux demandes sociales en matière d'évaluation des personnes. Une seconde version du Binet-Simon est publiée en 1908 et l'échelle est adaptée aux États-Unis dès 1909. Lewis Terman (1977-1956), professeur à l'université de Stanford, produit en 1916 le Stanford-Binet et l'épreuve fait ensuite l'objet de nombreuses adaptations. Durant la première guerre mondiale (1914-1918), Arthur S. Otis (1886-1964), élève de Terman, s'inspirera du Binet-Simon pour produire, à la demande de l'armée américaine deux tests collectifs utilisables pour la sélection et l'orientation des recrues : l'Army alpha (niveau normal) et l'Army Beta (niveau illettré). Grâce à ces possibilités de passation collective, 1,7 million de recrues ont été testés entre 1916 et 1918.

En 1912, Stern prolonge l'idée d'age mental de Binet en inventant un nouvel indice appelé quotient intellectuel (QI), rapport entre l'âge mental et l'âge chronologique. Il propose ainsi un indice de vitesse de développement intellectuel, interprétable en termes d'avance ou de retard. Cet indice sera très utilisé, et pas toujours à bon escient, et donnera lieu à de nombreuses controverses au XX^{e} siècle (voir Gould, 1983 ; Tort, 1974, Huteau et Lautrey, 1975) et encore actuellement (Lautrey, 2007).

L'approche factorielle de l'intelligence

➤ Un facteur général unique ?

À peu près à la même période où Binet et Simon travaillaient en France à leur échelle d'intelligence, Charles Spearman (1863-1945), chercheur londonien, envisage une autre approche de l'intelligence. Il est lui aussi élève de Wundt et est influencé par Galton et par ses découvertes statistiques. Il va en particulier perfectionner la mesure des corrélations et inventer l'analyse factorielle. Il pense que l'on peut s'appuyer sur cette analyse mathématique des performances des individus pour identifier les dimensions intellectuelles sur lesquelles les individus peuvent être comparés. Il publie en 1904 un article intitulé « General intelligence, objectively determined and mesured », dans lequel il expose les principes de sa méthode et les premiers éléments de sa théorie du facteur général d'intelligence. Après avoir fait passer différentes tâches très variées, essentiellement scolaires, à un échantillon de sujets et analysé les notes obtenues avec sa méthode de calcul, il obtient un facteur de variation commun à l'ensemble des épreuves et un facteur spécifique à chaque épreuve. Il appelle ce facteur commun facteur général d'intelligence ou facteur g. Son modèle factoriel de l'intelligence est précisé dans un ouvrage publié en 1927 (« The abilities of man, their nature and measurement »). Pour Spearman, le facteur g révélé par l'analyse factorielle correspond à de « l'énergie mentale ». Spearman met en évidence que les tâches les plus fortement saturées en facteur g sont des tâches « d'éduction de relations et de corrélats », c'est-à-dire d'extraction et d'applications de règles. Le facteur g reflète donc une capacité très générale à établir et appliquer des relations.

> Éduction de relations
> Consiste à trouver des relations entre plusieurs éléments.
> *Ex : Qu'y a-t-il de commun entre une voiture et un avion ?*
>
> Éduction de corrélats
> Consiste à trouver un objet à partir d'un autre, lorsque l'on connaît les relations qui les unissent.
> *Ex : Truite est à pêcheur ce que lapin est à... ?*

Ce modèle factoriel, appelé aussi « monarchique », est donc un modèle en deux niveaux de facteurs (modèle bi-factoriel) : le premier niveau correspond aux facteurs spécifiques à chaque tâche et le second niveau est celui du facteur commun à l'ensemble des tâches (facteur g). Ce modèle suppose donc que les tâches possèdent une part de variance commune. La réussite dans l'une,

covarie avec la réussite dans les autres. La figure 1.1. fournit une illustration schématique de ce modèle.

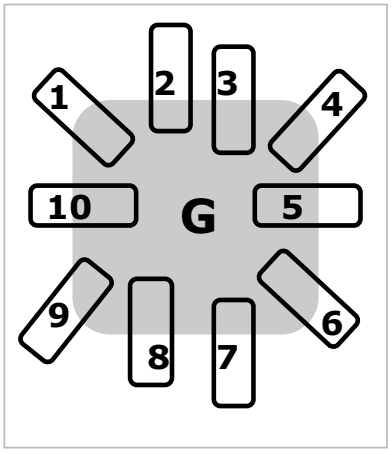

Figure 1.1
Représentation schématique du modèle de Spearman : les différentes épreuves (numérotées de 1 à 10) saturent pour partie dans un facteur unique appelé « facteur g ».

On peut noter que, bien qu'ayant adopté des approches méthodologiques et théoriques très différentes, Binet et Spearman partagent une conception globale et unidimensionnelle de l'intelligence. Cette position se retrouvera également dans l'approche de Daniel Wechsler (1896-1981) qui, à partir de 1939 proposera plusieurs échelles composites de mesure de l'intelligence. Weschler propose en 1939 une alternative au Binet Simon. Il adopte une nouvelle méthode de questionnement et une autre façon de calculer le QI que celle proposée par Stern (voir chapitre 3).

Il existe plusieurs tests qui ont été conçus en référence aux travaux de Spearman. Nous en présentons un certain nombre dans le chapitre 4 de cet ouvrage. En particulier John Raven s'inspirera de ces travaux de Spearman pour créer une épreuve fortement saturée en facteur g : l'épreuve des matrices (SPM).

On peut également noter que le facteur g et le QI sont tout deux des indices d'une intelligence unidimensionnelle, qui, bien qu'obtenus de façon très différente sont sur le fond très proches. De fait, les résultats aux Matrices de Raven corrèlent en moyenne à .80 avec des scores de QI (échelles de Weschler par exemple).

➤ Des aptitudes primaires indépendantes ?

Il est important de noter que les résultats de Spearman ont été obtenus en éliminant soigneusement, dans les épreuves choisies, toutes celles pouvant faire double emploi. Il a bien noté que lorsqu'il maintenait par exemple plusieurs tests verbaux dans sa batterie d'épreuve, ceux-ci saturaient certes dans le facteur G mais corrélaient également entre eux. Les tests verbaux saturaient donc également dans un facteur appelé « facteur de groupe » propre aux tests verbaux. Spearman a minimisé l'importance de ces facteurs de groupe. Au contraire, dans les années trente, l'accent a été mis sur l'existence et l'importance de ces facteurs de groupes par différents psychologues américains et en particulier par Louis, L. Thurstone (1887-1955). Thurstone a utilisé les techniques d'analyse factorielles, qu'il a contribué à perfectionner, et a constaté que lorsque l'on ne sélectionne pas comme le faisait Spearman de façon systématique les épreuves prises en compte, des facteurs de groupe apparaissent. Le facteur g serait donc le produit d'une sélection des épreuves et sa mise en évidence artificielle. Les facteurs que Thurstone identifie, et qu'il appellera « facteurs primaires » ont été par la suite retrouvés par la plupart des auteurs. Dans un ouvrage de 1935 intitulé *The vectors of the mind* il présente l'ensemble de son modèle et établit une liste de 9 facteurs primaires. Dans la suite de ses travaux, il en retiendra principalement 7 (voir tableau 1.2.). Pour Thurstone, ces facteurs sont indépendants et correspondent à des « aptitudes primaires » ou capacités intellectuelles qui structurent l'intelligence. Pour cet auteur l'intelligence n'est donc pas unidimensionnelle mais multidimensionnelle (ou multifactorielle), chaque aptitude primaire constituant une forme particulière d'intelligence.

De nombreux tests d'intelligence vont être inspirés de ce modèle. Tout d'abord, en 1938 paraît une première version de la batterie factorielle *Primary Mental Abilities (PMA)* mise au point par Thurstone lui-même. Cette épreuve est encore disponible et utilisée aujourd'hui. D'autres batteries factorielles sont présentées dans le chapitre 5.

➤ Peut-on concilier facteur G et aptitudes primaires ?

L'opposition entre le modèle de Spearman et celui de Thurstone n'est en réalité qu'apparente. En effet, dans le modèle de Thurstone, bien que les aptitudes soient présentées comme indépendantes les unes des autres, les recherches indiquent que des corrélations existent entre les facteurs primaires.

Tableau 1.2
Les sept aptitudes primaires (capacités intellectuelles) identifiées par Thurstone (1941).

	Capacités	Définition
N	Aptitude numérique	Rapidité et précision dans le traitement d'information chiffrée.
V	Compréhension verbale	Compréhension du langage et du vocabulaire.
W	Fluidité verbale	Production d'informations langagières nombreuses et variées.
S	Visualisation spatiale	Représentation mentale et traitement des objets, des lieux, des propriétés géométriques.
M	Mémorisation	Stockage et restitution d'informations.
R	Raisonnement inférentiel	Résoudre des problèmes par raisonnement logique (identifier les règles, appliquer les règles, faire des hypothèses).
P	Vitesse perceptive	Repérer rapidement des similitudes ou des différences sur des unités d'information élémentaires.

Thurstone n'a pas accordé beaucoup d'importance à ce fait, mais d'autres auteurs sont venus ensuite proposer des modèles plus complets qui vont réconcilier les points de vue de Spearman et de Thurstone dans des modèles « hiérarchiques » de la structure factorielle de l'intelligence.

C'est le cas de Burt et Vernon (*cf.* Vernon, 1950, 1952) ou encore de Cattel et Horn (Horn et Cattel, 1966, Cattel, 1971).

Ces auteurs analysent non seulement les saturations des tests dans les facteurs de groupe mais aussi les corrélations entre facteurs de groupes.

Cattel et Horn, deux psychologues qui travaillent aux États-Unis, procèdent par analyses factorielles successives. Après avoir extrait la variance expliquée par les facteurs primaires de Thurstone, ils mènent une seconde analyse factorielle (dite de second ordre) visant à extraire la variance commune aux facteurs primaires. Ils obtiennent ainsi plusieurs facteurs généraux, dont les 3 principaux sont : un facteur général d'intelligence fluide, un facteur général d'intelligence cristallisée et un facteur général d'intelligence visuo-spatiale. Les aptitudes de ces registres ont des propriétés distinctes. Celles qui relèvent de *l'intelligence fluide* sont, selon les auteurs, des produits de l'équipement neurologique et des apprentissages incidents. Ils conditionnent la réussite dans les activités qui impliquent la manipulation de relations complexes, la formation de concepts, le raisonnement et la

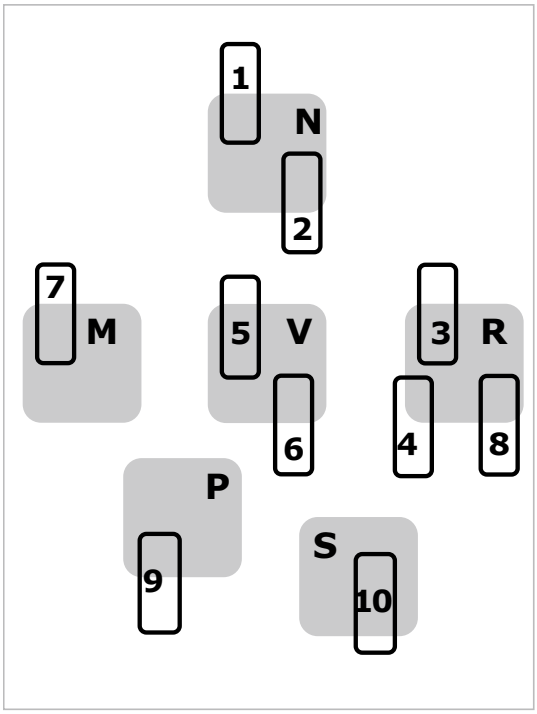

Figure 1.2
Représentation schématique du modèle de Thurstone : les différentes épreuves (numérotées de 1 à 10) saturent pour partie dans des facteurs de groupe appelés « facteurs primaires » qui correspondent à des registres intellectuels distincts.

résolution de problèmes nouveaux. Celles qui relèvent de l'*intelligence cristallisée* dépendraient de la culture, de la pratique scolaire, des apprentissages intentionnels, des habitudes, de l'expérience. L'intelligence fluide serait donc plus fortement déterminée par l'hérédité que l'intelligence cristallisée.

Cette distinction qu'ils introduisent entre intelligence fluide et cristallisée, sera ensuite reprise par de nombreux auteurs (*cf.* par exemple Baltes & Baltes, 1990).

Burt et Vernon, psychologues travaillant à Londres, procèdent également par des analyses factorielles mais optent pour une méthodologie différente. Alors que Cattel et Horn ont procédé à l'analyse des données du bas vers le haut, eux vont aller du haut vers le bas. Ils extraient tout d'abord la variance du facteur général dans la batterie de test utilisée, puis observent que la variance restante se partage entre deux grands facteurs de groupe. Ils appellent le premier Verbal-Education (V-E) et le second Kinesthésique-Moteur

(K-M). Une fois retirée la variance de ces deux grands facteurs, la variance restante se répartit entre plusieurs facteurs plus spécifiques qui correspondent aux facteurs primaires de Thurstone (*cf.* figure 1.3).

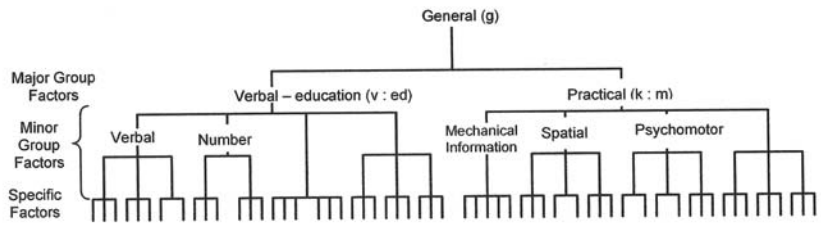

Figure 1.3

Représentation schématique du modèle hiérarchique de Burt et Vernon.

Les deux modèles hiérarchiques de ces auteurs présentent des points communs mais aussi plusieurs différences qui sont restées longtemps non résolues. Le modèle de Burt et Vernon comprend trois niveaux alors que celui de Cattel et Horn n'en contient que deux : il ne fait pas figurer de facteur général coiffant l'ensemble. Ce n'est que plus récemment que Gustaffson (1984) a apporté les éléments permettant d'expliquer les contradictions entre ces deux modèles. Grâce aux possibilités offertes par les analyses factorielles confirmatoires, il montre que le modèle le mieux ajusté aux données d'une batterie de tests (analyse factorielle restrictive) est bien un modèle en 3 niveaux comportant un facteur général. Il montre aussi que le facteur général de Burt et Vernon explique la même part de variance que le facteur d'intelligence fluide (Gf) de Cattel et Horn. Il parvient donc à un modèle stabilisé de la structure factorielle de l'intelligence, modèle qui sera confirmé et affiné quelques années plus tard par Carroll (1993). Nous présentons le modèle de Carroll, qui constitue la meilleure référence à ce jour sur la question, dans la partie suivante.

3. Principaux repères actuels de la psychométrie de l'intelligence

La structure factorielle de l'intelligence : modèles de synthèse

John B. Carroll, procède par méta-analyse, c'est-à-dire qu'il collecte les données issues de nombreuses études publiées relatives à la structure factorielle des tests d'intelligence et retraite ces données afin de trouver le modèle hiérarchique qui reflèterait le mieux l'ensemble de ces données. Il obtient les informations relatives à 460 études (ce qui correspond à plus de 130 000 sujets !) et constate que le modèle qui rend le mieux compte de ces données est un modèle en 3 niveaux (3 strates) qui intègre l'ensemble des modèles précédemment fournis. On y retrouve ainsi le facteur g de Spearman, ainsi que la structure multifactorielle de Thurstone et une structure hiérarchique qui concilie à la fois le modèle de Cattel et Horn et celui de Burt et Vernon. Ce modèle présente bien des similitudes avec celui de Gustaffson, mais offre une vision plus exhaustive et détaillée de la structure hiérarchique. Ce modèle, dont l'organisation est présentée dans la figure 1.4, fait aujourd'hui l'objet d'un large consensus.

La strate I correspond aux facteurs spécifiques (une trentaine), la strate II aux facteurs de groupe (8), et la strate III au facteur général. Les facteurs de la strate II sont hiérarchisés en fonction de leur niveau de saturation dans le facteur g. Ainsi, par exemple, les tests d'intelligence fluide sont de meilleurs représentants de l'intelligence générale que les tests de vitesse de traitement.

Nous pouvons noter que ce modèle distingue des formes variées d'intelligence (identifiées en particulier par les facteurs de la strate II). Ils présentent entre eux une relative indépendance qui rend compte du fait qu'un individu peut être performant dans un domaine sans nécessairement l'être dans tous les autres. Cela permet d'analyser les différences individuelles autrement qu'à travers un score unique sur une échelle unique et justifie le recours aux batteries factorielles d'intelligence. Néanmoins, la présence dans le modèle d'un facteur général témoigne d'une tendance statistique non nulle à ce que les résultats obtenus dans l'ensemble des épreuves corrèlent ente eux, ce qui donne également un sens à l'utilisation des épreuves de facteur g.

Les modèles multifactoriels hiérarchiques peuvent fournir une aide précieuse au praticien de l'évaluation. La sélection des tests ou des tâches à utiliser pour mener à bien une évaluation peut être éclairée par un

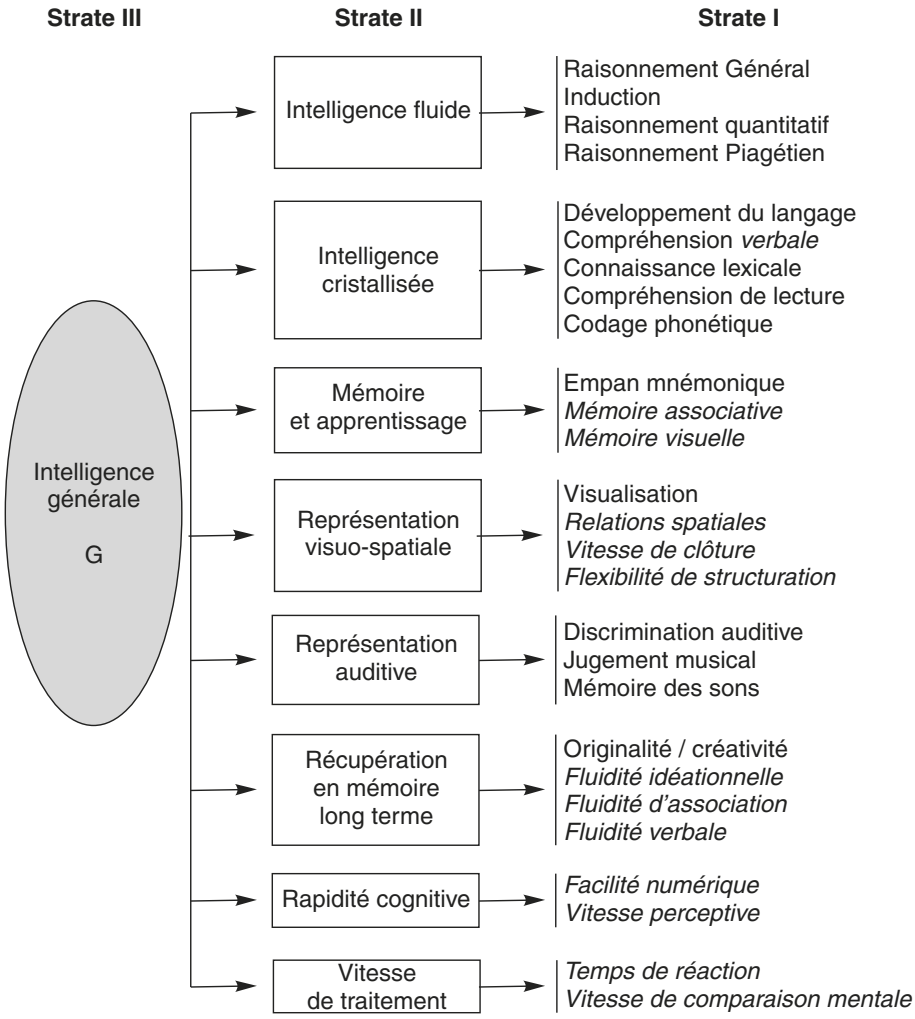

Figure 1.4
Structure hiérarchique des capacités cognitives (d'après Caroll 1993. Facteurs de vitesse en italiques et de puissance en caractères normaux).

positionnement des épreuves existantes, ou des types de tâches, en fonction de la place qu'occupent les capacités correspondantes dans le modèle hiérarchique (identifiée par l'analyse des saturations des items de ces tests dans les différents facteurs).

Nous présentons dans la figure 1.5 une cartographie de tests d'intelligence proposée par Snow, Kyllonen et Marshalek (1984), Snow et Lohman (1989) qui s'appuie sur une représentation en « Radex » élaborée à partir des travaux de Guttman (1957, 1965).

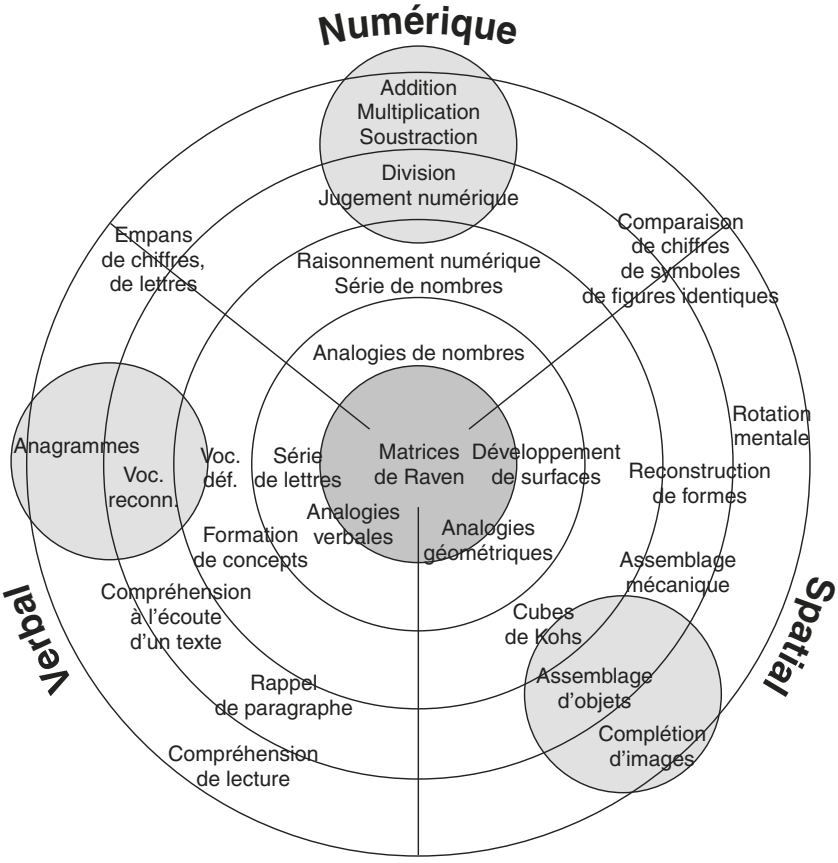

Figure 1.5
Représentation selon le modèle du Radex d'un ensemble fini de tests d'intelligence (d'après Snow et Lohman, 1989 et Lohman, 2000, cité par Juhel, 2005).

La figure représentée doit se voir comme un cône divisé en 3 grandes parties. Le radex fournit des informations sur la nature de ce qui est évalué et sur le niveau de simplicité ou de complexité cognitive des tâches. Il se lit de la façon suivante :

- Plus le test est proche du sommet du cône (centre de la figure), mieux il mesure le facteur *g ;*

- À l'inverse, plus un test est situé vers la périphérie, mieux il mesure des aptitudes spécifiques ;
- La nature des aptitudes évaluées dépend de la zone où est située l'épreuve. Une première région concerne les épreuves offrant un contenu figuratif ou dont les items sont des figures géométriques (domaine spatial), une seconde région rassemble les épreuves du domaine verbal (compréhension, lecture, vocabulaire...) et la troisième région correspond aux épreuves à contenu numérique (tâches impliquant des chiffres, des nombres, des calculs...).

La position du test dans l'espace de la figure informe donc sur la nature de la dimension qu'il évalue.

La position du test informe également sur le degré de complexité cognitive des épreuves (*cf.* Guttman et Lévy, 1991) :

- Les tests situés vers le sommet, requièrent de la puissance de raisonnement et sont intellectuellement plus complexes ;
- Le niveau intermédiaire marque des tâches plus simples où il est simplement nécessaire d'appliquer des règles sans avoir à les découvrir ;
- La base du cône correspond à des tâches plus spécifiques, s'acquérant principalement par apprentissages et pour lesquels la vitesse de réalisation est généralement importante.

Il est possible de choisir les différents subtests constitutifs d'une batterie en fonction de leur position dans l'espace du Radex. Cela constitue alors un élément supplémentaire de validité de l'épreuve par le choix de tâches non redondantes et couvrant plus largement l'ensemble des domaines et des niveaux d'évaluation. On peut également le faire à titre confirmatoire sur une batterie déjà existante. Une démarche de validation de ce type a été menée pour la batterie NV5R que nous présentons dans le chapitre 5.

Le niveau intellectuel est-il stable d'une génération à l'autre ?

Le psychologue néo-zélandais James R. Flynn a édité en 1984 et 1987 deux premières publications faisant état d'un phénomène peu pris en compte jusqu'à ce jour : une tendance à l'augmentation des résultats moyens dans les tests d'intelligence au fil des décennies et des générations. Ce phénomène a des conséquences importantes au plan théorique mais aussi au plan des pratiques de l'évaluation et mérite que l'on s'y attarde. Une très bonne synthèse sur le sujet a été publiée par Flieller en 2001.

➤ Description de l'effet Flynn

Flynn a constaté que « des cohortes de naissance testées au même âge et dans les mêmes conditions à l'aide d'une même épreuve d'intelligence obtiennent des scores moyens qui s'ordonnent comme leur année de naissance » (Flieller, 2001, p. 43).

Flynn a mené des travaux dans quatorze pays situés sur 4 continents et aboutit à la conclusion que la progression moyenne est d'environ 5 points de QI par décade, c'est-à-dire un écart-type par génération. D'autres études, réalisées depuis par divers auteurs un peu partout dans le monde, parviennent aux mêmes conclusions (voir par exemple Raven, 2001). L'effet Flynn, même s'il fluctue selon les pays, les périodes et les tests considérés, s'avère néanmoins un phénomène très général. Curieusement, la progression est plus importante dans les tests d'intelligence fluide que dans les tests d'intelligence cristallisée, ce qui paraît surprenant, puisque le registre de l'intelligence cristallisée est a priori mieux à même de profiter des apprentissages. En France, le phénomène a également été observé par plusieurs études : Flieller *et al.* (1986) observent une augmentation de 24 points de QI sur une période de 40 ans ou encore Baudelot et Establet, analysant les résultats des tests passés par les conscrits lors de leur incorporation, constatent une progression moyenne de 5 points de QI entre 1968 et 1982 (sur 14 ans).

En outre, des données anciennes attestent de la présence de ce phénomène dès la fin de la première guerre mondiale et montrent qu'il s'est prolongé à un rythme très régulier jusqu'à nos jours, concernant tous les âges de la vie (groupes de jeunes, d'adultes mais aussi de personnes âgées).

La grande généralité du phénomène est donc attestée et ne laisse pas d'interroger.

Tentatives d'explications du phénomène

Les tentatives d'explications sont nombreuses mais l'on doit bien admettre, comme le rappelle Flieller (2001), que le phénomène demeure encore une énigme.

Plusieurs hypothèses sont candidates à l'explication de l'effet Flynn. On retiendra en particulier :

- l'augmentation du brassage génétique des populations ;
- l'amélioration de la nutrition et des conditions d'hygiène et de santé ;
- l'augmentation des exigences et sollicitations cognitives de l'environnement ;
- les progrès de l'éducation.

Chacune de ces hypothèses est plausible. Cependant, les tentatives de validations menées pour chacune d'entre elles n'ont pas permis d'aboutir à des confirmations satisfaisantes. Le problème de l'origine de l'effet Flynn demeure donc entier.

Plusieurs pistes sont envisagées pour élucider le mystère. Certains privilégient l'hypothèse d'une origine plurifactorielle : plusieurs facteurs contribueraient conjointement à l'explication du phénomène. D'autres auteurs préfèrent remettre en cause l'interprétation du phénomène comme témoignant d'une augmentation du niveau intellectuel des populations : l'effet observé serait un artéfact du en particulier à la banalisation des tests (familiarisation des individus vis-à-vis des tests) ou encore à un changement d'attitude face aux situations d'évaluation par les tests. D'autres encore, et Flynn lui-même en fait partie, s'appuient sur la difficulté rencontrée à identifier les facteurs explicatifs du phénomène pour remettre plus fondamentalement en cause la capacité même des tests à évaluer l'intelligence.

Conséquences pour l'évaluation de l'intelligence

Quelles que soient les origines effectives de ce phénomène, celui-ci a des conséquences importantes d'une part pour notre connaissance de l'intelligence et d'autre part pour la qualité de sa mesure.

Concernant le premier point, l'effet Flynn, complique fortement les études développementales de l'intelligence et en particulier celles qui concernent le vieillissement cognitif. En effet, chaque fois que l'on est amené à étudier le développement en comparant des groupes d'âges différents (approche transversale), il devient difficile de savoir si les groupes sont réellement comparables et dans quelle mesure les résultats obtenus renseignent sur les effets de l'âge et ne sont pas dus à cet effet de cohorte. Nous reviendrons sur ce point dans le prochain paragraphe.

Concernant le second point, l'augmentation moyenne régulière des résultats aux tests d'intelligence accélère l'obsolescence des étalonnages des tests. On remarquera que, dans ce cas précis, le risque est, fort heureusement, de surestimer les résultats des individus dans les tests, et non l'inverse. Néanmoins, la validité de la mesure s'en trouve affaiblie et il est donc indispensable pour les auteurs et les éditeurs de réétalonner très régulièrement les tests et, pour le praticien de se garder d'utiliser des tests dont les étalonnages ne seraient pas récents (inférieurs à 10 ans). Les réétalonnages successifs, outre le coût qu'ils représentent, risquent de poser des problèmes méthodologiques relatifs au pouvoir discriminant des tests (lorsque le test devient par exemple trop facile pour tous). Nous évoquerons ces difficultés méthodologiques dans le chapitre 2 de cet ouvrage.

Le niveau intellectuel est-il stable chez l'adulte ?

Les premiers travaux portant sur l'évolution de l'intelligence chez l'adulte sont apparus avec le développement de la méthode des tests. Ils ont tout d'abord été menés principalement à l'occasion des étalonnages, c'est-à-dire selon la méthode transversale. La méthode consiste donc à comparer les performances de groupes de sujets d'âges différents et à inférer le développement de l'intelligence durant la vie à partir des performances moyennes obtenues aux différents âges (par ex. Jones & Conrad, 1933 ; Miles & Miles, 1932, Wechsler, 1939). Les données fournies par ces études ont étayé un modèle du développement de l'intelligence de l'adulte se présentant sous la forme d'un accroissement des capacités intellectuelles jusqu'à environ 20 ans, âge à partir duquel débute un déclin régulier qui s'accélère ensuite vers 60 ans.

C'est en se référant à ce modèle que certains auteurs ont pu émettre des doutes quant à la plasticité de l'intelligence de l'adulte de plus de 20 ans. C'est également ce modèle qui alimente encore très largement la représentation que le grand public a de l'intelligence de l'adulte.

On sait cependant aujourd'hui que les études transversales sont affectées par un biais méthodologique résultant de l'effet Flynn. Dans ces études, en effet, on ne compare pas seulement des sujets d'âges différents mais aussi des sujets de générations différentes. Les résultats caractérisant les différents âges n'étant pas obtenus sur les mêmes sujets, l'effet attribué à l'âge peut en réalité être dû, pour tout ou partie, aux différences de conditions de vie (éducation, santé, activités, stimulations) entre générations. En raison de ce biais, le modèle de l'évolution de l'intelligence issu de l'approche transversale, a été l'objet de nombreuses controverses (*cf.* Botwinick, 1977) et remplacé par un autre, issu d'études longitudinales.

Un vaste courant de recherche s'est développé à partir des années soixante-dix aux États-Unis (Birren, Dixon, Schaie, Willis...) et en Europe (notamment en Allemagne : Baltes et coll.) et a contribué, par des études longitudinales, à renouveler les connaissances sur l'intelligence de l'adulte. L'approche adoptée y est celle d'un « développement tout au long de la vie » (« *life span development* ») en rupture avec l'idée d'un développement s'arrêtant à la fin de l'adolescence et d'un âge adulte principalement marqué par le déclin.

Dans leur forme la plus simple, ces études longitudinales consistent en un suivi des mêmes sujets sur une certaine période, à l'aide d'évaluations répétées. Mais une difficulté demeure alors puisque le contrôle de l'effet

de cohorte n'est effectif que pour une seule génération. Pour pallier cette difficulté, les études visant l'obtention d'une vue d'ensemble de l'évolution « *life-span* » de l'intelligence ont eu recours à un plan plus sophistiqué (appelé séquentiel) qui est une combinaison des plans transversaux et longitudinaux. L'étude longitudinale est alors menée simultanément sur plusieurs cohortes, ce qui permet d'isoler les effets de cohorte des effets propres du vieillissement.

Nous pouvons retenir de ces travaux trois grandes conclusions :

1. Le déclin s'avère généralement plus tardif que ne le laissaient penser les études transversales

La plus importante étude longitudinale a été menée sous la responsabilité de Schaie (1979, 1983, 1994) : c'est l'« Étude longitudinale de Seattle ». L'étude a débuté en 1956 sur un échantillon de 500 sujets adultes âgés de 20 à 70 ans évalués à l'aide de différents tests d'intelligence dont les PMA de Thurstone. Ensuite, tous les sept ans, les auteurs ont procédé à la constitution d'un nouvel échantillon similaire et à l'évaluation des échantillons existants. Les dernières évaluations ont porté sur 8 cohortes de sujets âgés de 22 à 95 ans et, au total, près de 5000 personnes ont participé à l'étude. L'étude de Schaie montre ainsi que les performances dans les PMA de Thurstone ne commencent en moyenne à décroître qu'entre 50 et 60 ans (Schaie, 1994) (voir figure 1.6).

2. Le déclin n'affecte pas de la même façon les différents registres d'activité cognitive

Déjà dans les années soixante, Cattel et Horn avaient signalé une évolution différente avec l'âge des capacités relevant de l'*intelligence fluide et de l'intelligence cristallisée* (Horn, & Cattel, 1966) : les premières ayant tendance à décliner et les autres à se maintenir, voir à continuer de croître progressivement.

Ces différences ont été confirmées par un grand nombre d'études. Par exemple, Fontaine (1999) publie un tableau (voir tableau 1.3) issu des travaux de McGhee (1993) qui précise pour 9 grands domaines de capacités cognitives leur sensibilité aux effets négatifs du vieillissement (voir tableau 1.3).

Des évolutions dans la structure factorielle des aptitudes ont également été décrites. Symétriquement au processus de différenciation des aptitudes qui se manifeste dans l'enfance et à l'adolescence (Larcebeau, 1967 ; Nguyen-Xuan, 1969), un phénomène de dédifférenciation est observé en relation avec le vieillissement. Il se traduit par une diminution du poids des facteurs primaires et par une augmentation du poids du facteur g. Ce phénomène

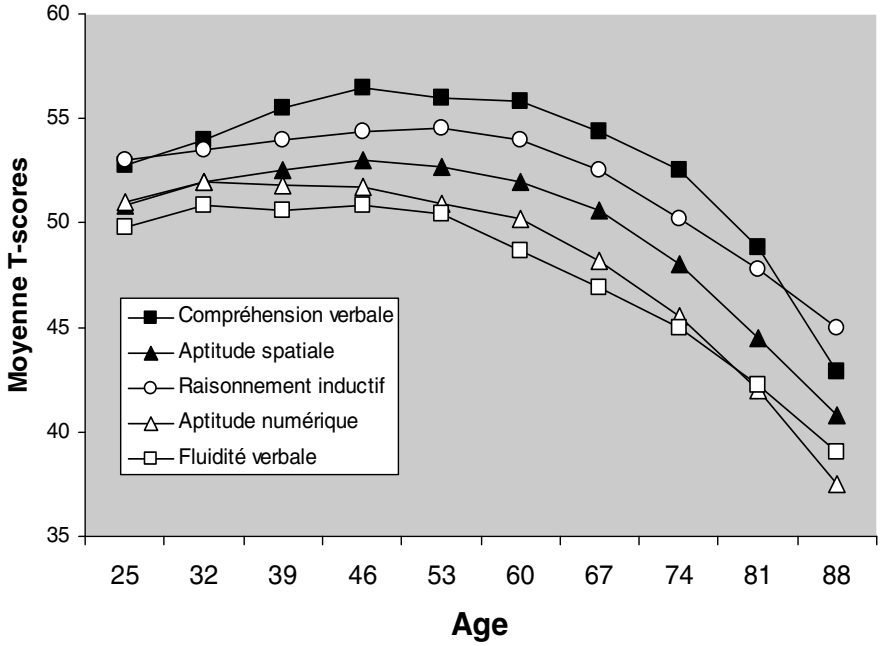

Figure 1.6
Courbes moyennes d'évolution avec l'âge des résultats dans les PMA de Thurstone observées dans l'étude longitudinale de Seattle (d'après Schaie, 1994).

initialement décrit par Balinsky (1941) a été confirmé et précisé par plusieurs études (Poitrenaud, 1972, Baltes *et al.,* 1980). Balinsky (cité par Fontaine, 1999) avait comparé des groupes d'âges différents et observé une diminution progressive des corrélations entre les subtests de la WAIS de 9 à 30 ans, puis une augmentation progressive de ces corrélations de 30 à 60 ans. Poitrenaud (1972) a observé une telle différence de structure factorielle entre deux groupes de sujets âgés respectivement de 64-69 ans et de 74-79 ans, alors que Lindenberger et Baltes (1997), comparant deux groupes âgés respectivement de 70-84 ans et de 85-103 ans, ne l'observent pas. On peut donc penser que cette dédifférenciation débuterait vers 30 ans et serait achevée vers 75 ans. Ce phénomène reste cependant controversé dans la mesure où il a principalement été observé par des études transversales et n'a pas trouvé de confirmation dans l'étude longitudinale conduite par Schaie. En outre, son étude présente un certain nombre de difficultés méthodologiques (Nesselroade et Thompson, 1995, Baltes *et al.* 1999).

Tableau 1.3
Tableau des domaines de capacités intellectuelles et de leur sensibilité au vieillissement (McGhee, 1993, Fontaine, 1999).

Nom	Définition	Sensibilité au vieillissement
Connaissance quantitative	Capacité à comprendre les concepts quantitatifs et leurs relations.	Faible
Compréhension, connaissance	Profondeur des connaissances.	Insensible
Mémoire à court terme	Capacité à enregistrer des informations et à les utiliser dans les secondes suivantes.	Sensible
Récupération à long terme	Capacité à enregistrer des informations et à les récupérer après un délai supérieur à quelques secondes.	Sensible
Processus auditif	Capacité à analyser et à synthétiser des stimulis auditifs.	Sensible
Vitesse de décision correcte	Capacité à répondre à des questions portant sur des problèmes de difficulté modérée nécessitant raisonnement et compréhension.	Très sensible
Raisonnement fluide	Capacité à raisonner, à construire des concepts, à résoudre des problèmes dans des contextes nouveaux.	Très sensible
Processus visuel	Capacité à analyser et à synthétiser des stimulis visuels	Sensible
Processus de rapidité	Capacité à réaliser rapidement des tâches cognitives automatiques sous pression et à maintenir l'attention	Très sensible

3. Une grande variabilité inter individuelle apparaît dans la façon de vieillir intellectuellement

La dispersion des résultats dans les tests augment avec l'âge (Nelson & Annefer, D., 1992). Lorsque l'on analyse cette variabilité on constate que l'avancée en âge ne se traduit pas pour tous les individus par les mêmes effets : les capacités intellectuelles ne déclinent pas de la même façon chez tous, pas nécessairement dans le même ordre, pas nécessairement au même âge, et pas avec la même intensité.

L'augmentation des différences interindividuelles avec le vieillissement pose assez logiquement la question des facteurs susceptibles d'influencer, de façon différentielle, ces évolutions cognitives liées à l'âge.

Plusieurs sources de variation ont été identifiées comme pouvant, seules ou en combinaison, contribuer à expliquer cette hétérogénéité.

Les facteurs les plus fréquemment évoqués sont relatifs aux conditions de vie actuelles de la personne, telles que son état de santé (*cf.* Herzog *et al.*, 1978 ; Perlmutter et Nyquist, 1990), l'intensité de sa vie sociale (*cf.* Moritz, 1989), ou son état marital (*cf.* Rogers, 1990), mais aussi aux caractéristiques de la personne avant qu'elle ne vieillisse, telles que son niveau culturel, la longueur de sa scolarité, ses activités professionnelles, son niveau intellectuel, ses antécédents de santé... (*cf.* Craik et al., 1987 ; Schaie, 1987 ; Ska *et al.,* 1997). Ces dernières variables, que Schaie (1990) appelle « antécédents des différences interindividuelles » peuvent ainsi jouer le rôle de prédicteur de la qualité du vieillissement.

Depuis une quinzaine d'années, des travaux ont été conduits visant à mieux connaître ces facteurs et la façon dont ils agissent, afin de déterminer les conditions optimales d'un « vieillissement réussi » (« *successful aging* »). L'une des hypothèses retenues par ces auteurs est que l'activité menée dans tel ou tel domaine puisse venir atténuer, voire totalement préserver de, certains effets négatifs du vieillissement et contribuer ainsi à expliquer les différences inter-individuelles dans la façon de vieillir (*cf.* Marquié, 1996, Loarer, 2000).

Par ailleurs des travaux menés dans le cadre de la psychologie cognitive ont permis d'identifier un certain nombre de processus cognitifs élémentaires particulièrement sensibles aux effets du vieillissement. Il s'agit en particulier de l'attention, de la mémoire de travail, de l'inhibition cognitive et de la vitesse de traitement. Ce dernier facteur apparaît essentiel (Salthouse, 1994, 1996) : la vitesse de traitement diminuant avec l'âge, le ralentissement cognitif pourrait contribuer fortement à la diminution des performances avec l'âge dans un grand nombre de registres. Pour un approfondissement de ces aspects, voir Lemaire et Behrer (2005).

Une ou plusieurs intelligences ?

La question de l'unicité ou de la pluridimentionnalité de l'intelligence était déjà présente dans l'opposition entre Spearman et Thurstone. On pourrait penser qu'elle a été résolue par les modèles hiérarchiques synthétiques que

nous venons de présenter. On peut pourtant s'interroger sur le fait que ces modèles refléteraient la totalité de ce qui caractérise l'intelligence humaine.

En particulier, dès lors que l'on considère l'individu engagé dans des tâches et des situations pratiques de la vie quotidienne, l'intelligence évaluée par les tests peut sembler insuffisante pour rendre compte de l'ensemble de ses fonctionnements adaptatifs.

L'interrogation n'est pas récente et de nombreux auteurs ont opté pour une vision élargie de l'intelligence. Déjà, en 1920, Edward L. Thorndike (1874-1949) identifiait 3 facettes à l'intelligence. Il la définissait comme l'habileté à comprendre et à gérer 1/ les idées (intelligence abstraite), 2/ les objets (intelligence concrète ou mécanique) et 3/ les personnes (intelligence sociale). Cette dernière facette est ainsi définie par Thorndike comme la capacité à « comprendre et gérer les autres personnes » et à « agir sagement dans les relations humaines » (1920, p. 228). L'intelligence classiquement évaluée par les tests d'intelligence ne correspond donc principalement qu'à la première de ces facettes.

Par ailleurs, Weschler s'est également un temps intéressé à ce qu'il appelait les « facteurs non-intellectifs » de l'intelligence (1943, p. 108). Il désignait par là en particulier l'intelligence sociale. Il conclura cependant quelques années plus tard que l'intelligence sociale n'est rien d'autre que de « l'intelligence générale appliquée aux situations sociales » (1958, p. 75).

Cette question a cependant repris de la vigueur dans la période récente. Elle correspond à la tendance de plus en plus affirmée à ne pas considérer l'intelligence uniquement sous l'angle de la pensée logicomathématique mais d'accorder une place plus importante aux différentes facettes des activités mentales qui contribuent à l'adaptation de l'individu et à son efficacité dans les différentes sphères de ses activités. Cette tendance s'exprime notamment dans le modèle de l'intelligence de Sternberg, en particulier par la prise en compte de formes d'intelligence dites « pratiques » ou « non académiques » (Sternberg, 1985, Sternberg et al., 2000), ainsi que dans le modèle des intelligences multiples de Gardner (1996, 1999) ou encore dans les travaux menés sur l'intelligence émotionnelle (Salovey & Mayer, 1990) et sur l'intelligence sociale (voir Loarer, 2005 sur ces deux aspects).

Ces modèles et ces travaux présentent à notre avis l'intérêt d'élargir la notion d'intelligence pour chercher à mieux prendre en compte l'étendue de la palette des ressources adaptative des individus et à mieux saisir ce qui sous-tend l'organisation des conduites dans les situations de la vie quotidienne. L'intelligence cognitive, celle qui prend appui sur le traitement logique de l'information, joue à l'évidence un rôle essentiel pour

permettre aux individus de développer des conduites « intelligentes », mais d'autres registres, notamment émotionnel et sensori-moteurs, y contribuent à l'évidence également et gagnent à être mieux étudiés, en particulier dans leurs interactions avec la cognition. Ils gagneraient également à être mieux évalués, mais actuellement, notamment en France, les tests dans ces domaines restent malheureusement encore peu nombreux.

Dans le cadre de cet ouvrage, nous avons fait le choix de nous centrer uniquement sur l'intelligence cognitive et son évaluation.

Points de repères clés dans l'approche psychométrique de l'intelligence	
1879	Wundt crée à Leipzig le premier Laboratoire de Psychologie Expérimentale.
1884	Galton applique à grande échelle des tests standardisés anthropométriques et sensoriels et réalise des étalonnages.
1890	Cattel propose le terme de « mental tests » pour des épreuves évaluant encore principalement les fonctions élémentaires.
1904	Spearman introduit l'analyse factorielle et la notion de « facteur g » (facteur général d'intelligence).
1905	Binet et Simon créent la première échelle de mesure de l'intelligence évaluant les fonctions supérieures et proposent la notion d'âge mental.
1912	Stern propose la notion de Quotient Intellectuel (QI).
1916	Terman adapte l'échelle de Binet et Simon aux États-Unis.
1917	Otis crée les premiers tests collectifs : *l'Army alpha* et *l'Army beta*.
1935	Thurstone publie *The vectors of the mind* qui décrit un modèle multifactoriel de l'intelligence.
1938	Raven publie la première version des Matrices Progressives inspirée des travaux de Spearman.
1939	Weschler crée le Weschler-Bellevue et propose une autre façon de calculer le QI.
1947	Bennet et coll. publient le *Differential Aptitude Test (DAT)*, batterie multifactorielle inspirée des travaux de Thurstone.
1952	Burt et Vernon proposent un modèle hiérarchique en 3 niveaux conciliant les positions de Sperman et de Thurstone.
1966	Cattel et Horn proposent un second modèle hiérarchique dans lequel ils distinguent l'intelligence fluide (Gf) et l'intelligence cristallisée (Gc).
1984	Gustaffson utilise l'analyse factorielle confirmatoire pour concilier les modèles de Cattel-Horn et de Burt-Vernon.
1985	Naglieri publie aux États-Unis le NNAT.
1983	Kaufman et Kaufman, publient le K-ABC, version américaine (version française en 1993).
1993	Caroll propose un modèle synthétique hiérarchique de l'intelligence fondé sur une vaste méta–analyse.
Ces quelques repères n'ont pas la prétention d'être exhaustifs mais correspondent aux principaux auteurs et épreuves cités dans cet ouvrage qui ont marqués l'histoire de l'évaluation de l'intelligence.	

CHAPITRE
2

Définition et propriétés des tests

Sommaire

DANS le chapitre 1 nous venons de présenter les éléments historiques et théoriques de la notion d'intelligence et de sa mesure. Nous allons maintenant aborder les aspects méthodologiques relatifs à cette mesure. En effet, dès le début du XX^e^ siècle apparaissent les premiers tests d'intelligence qui permettent d'évaluer, de mesurer et de rendre compte des différences interindividuelles dans ce domaine. Il va s'agir alors d'élaborer des dispositifs d'observation qui présentent un bon niveau de fiabilité. Les tests, et spécifiquement ici les tests d'intelligence, ne relèvent donc pas d'une approche « magique » comme nous le signalent Huteau et Lautrey mais peuvent être considérés comme des techniques d'observation :

> « Les tests ne révèlent pas des propriétés mystérieuses et cachées des individus. Ils permettent simplement de décrire des comportements sous l'angle de leur efficience. Ils ne sont rien d'autre que des techniques d'observation. » (Huteau et Lautrey, 1997, p. 3.)

Les tests sont donc des outils de mesure qui doivent présenter, comme tout instrument de mesure, certaines qualités métrologiques (appelées ici psychométriques), que nous allons détailler dans ce chapitre[1].

L'objectif principal ici est d'exposer les principaux éléments méthodologiques relatifs aux tests d'intelligence. Le lecteur intéressé par l'approfondissement du sujet pourra se référer à des ouvrages plus complets comme par exemple celui de Dickes *et al.* (1994) ou de Laveault et Grégoire (2002).

1. Définitions préalables

Qu'est-ce qu'un test ?

Il convient, pour commencer, de définir précisément ce qu'on entend par test psychologique. Parmi les nombreuses définitions proposées, prenons par exemple celle de Pichot :

> « On appelle test mental une situation expérimentale standardisée servant de stimulus à un comportement. Celui-ci est comparé statistiquement à celui d'autres individus placés dans la même situation, de manière à classer

1. Nous remercions Pierre Vrignaud pour sa lecture critique d'une première version de ce chapitre.

> le sujet examiné par rapport à ceux constituant le groupe de référence. » (Pichot, 1997, p. 5.)

Pour Pichot, un test correspond donc à un certain type de situation (une situation *expérimentale standardisée*), situation qui vise à produire chez le sujet un certain comportement. C'est ce comportement qui va être mesuré. Mais la mesure en elle-même n'a guère de sens, car c'est par la comparaison *statistique* de ce comportement à celui d'autres individus placés dans la même situation que cette mesure va acquérir une signification.

Cette notion de situation standardisée apparaît également dans la définition proposée par Huteau et Lautrey en 1997. Pour ces auteurs :

> Un test est « un dispositif d'observation des individus qui présente quatre propriétés :
> – il est standardisé ;
> – il permet de situer la conduite de chaque sujet dans un groupe de référence ;
> – le degré de précision des mesures qu'il permet est évalué (fidélité) ;
> – la signification théorique ou pratique de ces mesures est précisée (validité). » (Huteau & Lautrey, 1997, p. 19.)

Cette définition indique clairement les principales caractéristiques des tests avec les notions de standardisation, de groupe de référence, de fidélité et de validité. Nous allons développer et illustrer ces différentes notions mais nous pouvons déjà donner quelques indications sur ce qu'elles recouvrent :

- La *standardisation* est la définition précise des conditions de passation (matériel, consignes, temps...) et des conditions de cotation (modalités de correspondance entre une conduite et un score, calcul des scores...) ;
- Le *groupe de référence* est un groupe de sujets qui présentent les mêmes caractéristiques que le sujet examiné (au niveau de l'âge et/ou du sexe et/ou du niveau d'étude...) et qui a été confronté au même test. L'ensemble des scores obtenus par ce groupe de sujets va permettre de situer les performances d'un sujet examiné dans ce groupe de référence (principe de l'étalonnage) ;
- La *fidélité* correspond à l'une des qualités psychométriques attendues d'un instrument de mesure : la stabilité de la mesure. Nous verrons que l'on peut distinguer plusieurs types de fidélité ;
- La *validité* correspond à ce qui est mesuré par le test. Par exemple pour les tests dont il sera question dans cet ouvrage il faut s'assurer qu'ils évaluent

tous l'intelligence. Nous verrons également qu'il existe plusieurs types de validité.

Ces deux définitions de la notion de test indiquent bien, d'une part, qu'un test n'est pas un instrument magique et mystérieux (il s'agit d'un dispositif précis, explicite, visant à mesurer un comportement) et, d'autre part, qu'un test doit présenter certaines qualités (ce qui permet de distinguer les tests d'autres situations d'évaluation, comme par exemple les *questionnaires* de magazines, qui ne présentent pas ces caractéristiques...).

Comment se présente un test ?

Un test est constitué d'un ensemble de petites situations d'évaluation. Ces situations sont le plus souvent des questions auxquelles le sujet doit répondre, ou des petits problèmes auxquels il est confronté. Ces problèmes peuvent également prendre la forme de tâches à accomplir, comme par exemple dans le test des cubes de Kohs où le sujet doit reproduire une figure dessinée à l'aide de cubes colorés.

Chaque question ou chaque problème est appelé *item***.** Un test comporte ainsi plusieurs items, entre une vingtaine et une soixantaine selon les tests.

Le psychologue dispose généralement de plusieurs documents pour un même test :

- le test proprement dit, qui peut prendre par exemple la forme d'un cahier de passation où figurent les items,
- une feuille de réponse (ou protocole) ;
- Le manuel du test, qui comporte un ensemble d'informations relatives à la passation et à la cotation, mais également des informations sur l'élaboration et la validation de l'épreuve. Ce manuel peut éventuellement comporter des annexes éditées postérieurement au manuel, et destinées à le compléter.

Comment passer d'un comportement à un score ?

Pour chaque item, la performance du sujet est évaluée en fonction de la qualité de sa réponse (bonne ou mauvaise) ou de la qualité de la tâche réalisée. On accorde ainsi, le plus souvent, un point par bonne réponse. Le temps de résolution peut également être pris en compte (soit par item, soit sur

l'ensemble de l'épreuve). Au final, on fait la somme de l'ensemble des points obtenus par le sujet dans ce test pour obtenir un score, appelé *score brut*.

Ce score brut n'a pas de valeur en soi. Par exemple, si le test comporte 60 items et que le sujet obtient 43 points (donc 43 bonnes réponses), comment interpréter ce score brut de 43 ? Est-ce une bonne performance ? Sans informations sur le niveau de difficulté du test, et/ou sur le degré de réussite d'autres sujets confrontés à la même épreuve, il n'est pas possible d'interpréter ce score de 43 points. Pour cela, il faut utiliser un étalonnage qui va nous permettre de situer la performance du sujet par rapport aux résultats d'un groupe de sujets comparables au sujet examiné. C'est par cette comparaison que l'on pourra interpréter le niveau de performance du sujet. En reprenant notre exemple de 43 points, l'étalonnage va nous indiquer si ce score de 43 se situe en dessous ou au-dessus de la moyenne du groupe, ce qui est un élément de réponse. Les étalonnages étant en général détaillés, ils permettent de situer plus précisément le niveau de performance du sujet qu'en le référent simplement à la moyenne. Par exemple, si l'étalonnage utilisé nous indique que seulement 10 % des sujets obtiennent un score égal ou supérieur à 43 points, ce score brut de 43 reflétera alors un très bon niveau de performance. L'étalonnage permet donc de transformer une note brute en note étalonnée et ainsi de positionner précisément la performance du sujet au regard de celle d'un groupe de référence. Cela permet l'interprétation du résultat obtenu au test. Nous verrons plus loin (en 2.10) qu'il existe différents types d'étalonnages.

Les scores étalonnés permettent également de comparer les performances d'un même sujet face à des épreuves différentes lorsque la comparaison directe sur les scores bruts n'est pas possible : par exemple, un sujet confronté à deux tests qui comportent le même nombre d'items mais qui ne sont pas du même niveau de difficulté ou confronté à des tests qui ne comportent pas le même nombre d'items.

Les différents types de tests

Il existe de nombreux tests et l'on peut les classer selon différents critères : en fonction du domaine évalué, du mode de passation, du format ou du type de réponse.

Nous allons évoquer ces classifications en les illustrant par quelques exemples de tests d'intelligence.

➤ Classification des tests en fonction du domaine évalué

On peut distinguer les épreuves en fonction du domaine ou des dimensions qui sont évaluées : tests d'intelligence, tests de personnalité (intérêts, motivation, traits de personnalité), épreuves de créativité, tests de connaissances (connaissances scolaires et/ou connaissances relatives à un domaine précis comme par exemple la mécanique ou l'informatique...), tests psychomoteurs (tests d'efficience motrice, de latéralité...).

Au sein de chaque domaine, une catégorisation plus fine peut être effectuée. Ainsi, dans le domaine des tests d'intelligence, il est possible de distinguer les échelles d'intelligence, les tests de facteur g, les batteries factorielles, les tests verbaux... en lien direct avec les conceptions théoriques sous jacentes (voir chapitre 1).

➤ Classification des tests en fonction du format

Le format du test correspond au type de support utilisé ou au type de matériel de passation :

- Lorsque le sujet dispose d'un feuillet ou d'un livret de passation et doit indiquer sa réponse par écrit, on parle *de test « papier/crayon »* ;
- Si le sujet doit effectuer une tâche (comme par exemple reproduire une figure à l'aide de cubes colorés ou remettre en ordre des images) on parle alors de *test de performance* ;
- Enfin, lorsqu'un ordinateur est utilisé pour la passation, pour les questions et/ou pour les réponses, ou pour proposer des tâches à effectuer, il s'agit de *test informatisé*.

➤ Classification des tests en fonction du mode de passation : individuel ou collectif

Certains tests sont conçus pour être administrés en situation individuelle, comme par exemple les cubes de Kohs, les échelles de Wechsler... Dans ce cas, un seul sujet est face au psychologue. Tous les tests de performance, tests dans lesquels le sujet doit effectuer une tâche, sont des tests individuels car le psychologue ne peut observer qu'un seul sujet à la fois.

D'autres épreuves sont conçues pour une passation collective, ce sont généralement des tests « papier/crayon », comme par exemple les matrices de Raven. Dans ce cas, chaque sujet dispose d'un cahier de passation et d'une

feuille de réponse. Par cette procédure, plusieurs sujets peuvent être testés en même temps par un seul psychologue. Bien entendu, un test collectif peut toujours être utilisé en passation individuelle, l'inverse ne l'est pas forcément.

Le test individuel permet de recueillir davantage d'informations telles que le comportement du sujet au cours du test, et plus particulièrement ses réactions face à une difficulté, le non verbal, le repérage des erreurs... Le test collectif permet quant à lui un gain de temps aussi bien pour la passation que pour la cotation.

➤ Classification des tests en fonction du type de réponse

Dans certains tests, et plus fréquemment dans les tests d'intelligence, il n'existe qu'une seule bonne réponse à chaque item[1]. Mais cette catégorisation de type 0/1 (1 point en cas de bonne réponse, 0 point pour toute autre réponse) peut être affinée comme par exemple dans certains subtests des échelles de Wechsler dans lesquels, en fonction de la qualité de la réponse (spécificité ou généralité des propos...), une bonne réponse compte pour 1 point ou pour 2 points ; ou encore dans d'autres tests qui envisagent de retirer des points pour les mauvaises réponses (et tentent d'éviter ainsi les réponses au hasard).

On distingue les questions *ouvertes* des questions *fermées* :

- Par question *ouverte* il faut entendre réponse à construire, comme par exemple dans le test D2000 où le sujet doit créer sa réponse (voir la présentation de ce test dans le chapitre 4) ;
- Par question *fermée* il faut entendre Q.C.M (Question à Choix Multiples) dans lesquels le sujet doit choisir sa réponse parmi plusieurs possibilités, comme par exemple dans les matrices de Raven (voir la présentation de ce test dans le chapitre 4).

Les questions ouvertes présentent l'avantage de recueillir un maximum d'informations, les questions fermées présentent l'avantage d'une rapidité, et d'une fidélité, de la cotation. Mais il semble que les avantages et inconvénients relatifs de ces deux possibilités de réponse sont en réalité un peu plus complexes (voir Vrignaud, 2003).

1. Dans d'autres tests, la notion même de bonne réponse n'a pas de sens : tests de personnalité, questionnaire d'intérêts...

La notion de psychométrie

Le terme « psychométrie » évoque sans doute chez beaucoup de psychologues les tests, et uniquement les tests. On peut cependant considérer que cette notion concerne plus largement « l'ensemble des théories et des méthodes de la mesure en psychologie » (Dickes *et al.*, 1994, p. 11). La psychométrie dépasse ainsi l'analyse des tests et concerne plus largement toute approche psychologique visant à attribuer des nombres à des objets d'étude. D'ailleurs ces mêmes auteurs affirment, de façon sans doute un peu provocatrice, qu'il est possible de traiter de la psychométrie sans référence aux tests !

> « La psychométrie peut se concevoir sans aucune référence aux tests. L'étude des tests et de leur construction fait certes partie de la psychométrie, mais elle n'en est qu'une partie limitée : les tests ne sont qu'une classe d'instruments de mesure parmi d'autres. » (Dickes *et al.*, 1994, p. 11).

Dans ce cadre, comme nous l'illustrerons plus loin, différents modèles de mesure peuvent être utilisés.

Nous retiendrons ici que la psychométrie représente les théories et les méthodes qui permettent d'élaborer les instruments de mesure psychologique que sont les tests et d'en interpréter les résultats. Nous ne présenterons dans cet ouvrage que certains aspects de la psychométrie, ceux qui nous semblent être en lien direct avec notre propos. Les lecteurs intéressés par une présentation plus complète des différents aspects de la psychométrie peuvent consulter l'ouvrage de Dickes *et al.* (1994) ou celui de Laveault et Grégoire (2002).

Les théories et les méthodes psychométriques permettent d'élaborer des tests dans l'objectif de garantir une objectivité de la mesure. Il s'agit alors, aussi bien dans les phases d'élaboration et de validation des épreuves que lors de l'utilisation de ces tests, de s'assurer de la fiabilité de la mesure.

La psychométrie nécessite l'usage, et la compréhension, de quelques connaissances en statistiques et, en particulier, la notion de corrélation. En effet, les coefficients de corrélation sont très souvent utilisés comme indicateurs de la fiabilité d'un test.

Pour revoir ces notions statistiques le lecteur peut consulter des ouvrages de base (voir par exemple Guéguen, 2005 ; Corroyer et Wolff, 2003 ; Beaufils, 1996a et b).

Rappel sur la corrélation

La corrélation est un indicateur statistique qui permet de juger du degré de liaison existant entre deux séries d'observation. L'indicateur le plus connu est le coefficient « r » de Bravais-Pearson (appelé par la suite r de B-P) qui estime le sens et l'intensité du niveau de liaison linéaire existant entre deux variables quantitatives, comme par exemple la liaison entre les scores d'un même groupe de sujet confronté à deux tests. Cet indicateur r peut, par construction, prendre les valeurs comprises entre –1 et +1.

Rappelons qu'il s'agit ici, avec cet indicateur r de B-P, d'analyser une liaison *linéaire* et qu'il peut exister d'autres formes de liaison entre deux variables, comme par exemple une liaison quadratique...

Le sens de la liaison est indiqué par le signe du r de B-P : un r positif signifie que les deux variables varient dans le même sens, un r de B-P négatif signifie que les deux variables varient en sens inverse. Par exemple, si on calcule un r de B-P entre deux tests d'intelligence on s'attend à obtenir une valeur positive : les sujets ayant un score élevé dans un des tests devraient avoir un score élevé dans l'autre test, et inversement pour les scores faibles. Les deux variables varient bien dans le même sens car il s'agit de la même dimension : ces deux tests évaluant le même domaine.

L'intensité (la force) de la liaison est estimée par la valeur du r de B-P : plus le r de B-P est proche de 1, plus la liaison est forte ; plus il est proche de 0, plus la liaison est faible. Un r de B-P égal à 1 exprime donc une liaison parfaite (dans la réalité un r de 1 est quasiment impossible à observer), un r de B-P égal (ou proche) de 0 exprime une absence de liaison entre les deux variables. Par exemple, si on calcule un r de B-P entre deux tests d'intelligence, on doit observer une valeur plus proche de 1 que de 0 car les deux tests évaluent la même dimension : les sujets ayant un score élevé dans l'un des tests doivent également avoir un score élevé dans l'autre test.

Le sens et la force de la liaison peuvent aussi s'interpréter à partir d'un diagramme de corrélation : plus le diagramme des points est proche d'une ellipse étroite, plus la liaison est forte ; plus le diagramme des points est proche d'un cercle, plus la liaison est faible.

Comme nous le verrons par la suite, la corrélation, et principalement le r de B-P, est souvent utilisée pour qualifier les qualités psychométriques des épreuves (validité, fidélité...). Des valeurs sont alors indiquées dans les

notices des tests mais le problème important est celui de l'interprétation : comment interpréter ces valeurs ?

L'interprétation du r de B-P va tout d'abord dépendre de la situation. Par exemple, s'il s'agit de qualifier la relation entre deux passations d'une même épreuve sur les mêmes sujets (situation de la fidélité test-retest) on s'attend à une valeur de r très proche de 1 et au minimum de .80[1]. Par contre, lorsque l'on souhaite analyser la relation entre une épreuve et un critère, comme par exemple la réussite scolaire (situation d'analyse de la validité prédictive) on s'attend alors à des valeurs de r plus faibles, autour de .50, car on sait que d'autres variables que l'efficience intellectuelle ont des effets sur la réussite scolaire et que cette multiplicité de facteurs a comme effet de réduire le pouvoir explicatif d'une seule variable.

La valeur d'un coefficient de corrélation est donc toujours à interpréter en fonction de la situation. En l'absence de normes clairement définies nous indiquerons, à chaque fois que c'est possible, des valeurs repères qui permettront aux psychologues de juger des valeurs qu'il peut trouver dans les notices des tests (les manuels) ou dans des articles.

La standardisation

Les définitions du test que nous avons présentées dans notre introduction insistent sur la standardisation de la situation d'évaluation. On peut en effet dire que la standardisation conditionne la possibilité de comparaison des résultats. Dans une situation standardisée tout est soigneusement défini : de la présentation du test aux modalités de calcul des scores.

- Pour les *modalités de passation* : le matériel, les consignes, les temps de présentation et de résolution, les items exemples, l'attitude du psychologue, les éventuelles aides ou relance en cas d'échec, les éventuelles règles d'autocorrection de ses erreurs[2], les conditions d'arrêt de passation...
- Pour les *modalités de cotation* : les conditions d'attribution des points, les conduites à tenir face aux mauvaises réponses (parfois elles peuvent venir en déduction des scores), les modalités de calcul du ou des scores...

1. L'usage veut que pour les indices statistiques inférieurs à 1 (en valeur absolue) on utilise une notation de type .80 au lieu de 0,80. Nous adopterons donc cette notation car c'est celle que le lecteur trouvera par exemple dans les manuels de tests.
2. Comme par exemple la possibilité ou non de revenir en arrière afin de corriger une réponse...

Tous ces éléments, aussi bien pour la passation que pour la cotation, sont généralement clairement définis, voire illustrés par des exemples, dans le manuel du test. Ces informations permettent de placer tous les sujets *exactement* dans la même situation, et plus particulièrement, de les placer dans la même situation que les sujets de l'échantillon d'étalonnage. Si ces conditions sont équivalentes, la standardisation est assurée : on peut alors se reporter avec confiance à l'étalonnage pour situer le niveau de performance du sujet. C'est cette standardisation qui permet la comparabilité des résultats. Sinon, si l'un des éléments de standardisation est défectueux, la situation d'évaluation est différente de celle qui est prévue (par exemple, si on fournit des aides au sujet pendant la passation, si on lui laisse plus de temps...) et on ne peut plus utiliser l'étalonnage.

L'un des objectifs de la standardisation est d'éviter un certain nombre de biais potentiels, et plus particulièrement les biais de cotation relatifs à l'observateur. Par exemple, la standardisation permet de garantir une fidélité inter juge de la cotation : deux psychologues confrontés au même protocole doivent aboutir au même résultat (c'est-à-dire au même score). Dans les épreuves de type QCM cette fidélité devrait être parfaite. Dans les épreuves comportant des questions ouvertes, si le psychologue suit correctement les consignes de cotation, cette fidélité doit également être très bonne. Par exemple, dans le manuel du WISC-III, cette fidélité a été vérifiée par la cotation de 60 protocoles par deux psychologues. Les coefficients de fidélité inter cotateurs observés sont très élevés avec des valeurs autour de .99 pour les épreuves verbales (manuel WISC-III, p. 185).

C'est cette standardisation qui distingue les évaluations psychologiques d'autres évaluations, comme par exemple les évaluations scolaires dont on sait, depuis Piéron, qu'elles présentent un certain nombre de biais (Piéron, 1963). D'ailleurs, pour des évaluations scolaires il existe des tests normalisés de connaissances, de type Q.C.M, qui reposent sur les mêmes méthodologies psychométriques que les tests psychologiques (voir par exemple les tests T.A.S[1] diffusés par les ECPA) et qui garantissent ainsi un niveau de fiabilité plus élevé que les évaluations scolaires classiques (pour la comparaison entre l'évaluation scolaire et l'évaluation psychologique voir Huteau, 1996).

Le psychologue doit donc bien comprendre l'utilité du respect des consignes qui sont énoncées dans le manuel du test, aussi bien comme nous venons de le voir pour la passation que pour la cotation. Même si

1. Test d'Acquisition Scolaire.

ces consignes peuvent parfois sembler rigides, le praticien doit se garder de prendre quelques libertés avec celles-ci.

Le respect de la standardisation intervient, comme nous allons le voir, dans la fidélité d'un test.

2. La notion de fidélité

Le principe de fidélité

La fidélité de la mesure (on parle aussi de fiabilité) concerne la constance des résultats obtenus :

> « On considère qu'un instrument de mesure est fidèle si le résultat qu'il produit est reproductible. » (Dickes *et al.*, 1994, p. 165).

Cette notion est importante car elle conditionne la fiabilité du test : un test qui n'est pas fidèle ne peut être ni fiable, ni valide.

Une mesure est fidèle si, confrontés plusieurs fois à cette mesure, les sujets obtiennent des résultats comparables (en éliminant les possibles effets d'apprentissage entre les passations). La fidélité est une qualité de tout instrument de mesure : par exemple, une balance doit indiquer un même poids à quelques minutes d'intervalle, une toise doit indiquer une même taille, un mètre ruban doit indiquer une même longueur. Il en est de même pour les tests.

En restant dans le domaine de la psychométrie, l'exemple le plus simple est la notion de fidélité dans le temps. Si un même groupe de sujets passe le même test à quelques semaines d'intervalle on doit observer les mêmes résultats ou, au moins, des résultats comparables. Il s'agit ici de la fidélité, ou stabilité, temporelle par la méthode dite « test/retest ». On peut évaluer cette fidélité par le calcul d'un r de B-P entre les deux passations. Nous verrons qu'il existe plusieurs types de fidélité. La fidélité de la mesure contribue à la fiabilité de la mesure.

Mais cette fidélité n'est jamais parfaite. La mesure répétée à plusieurs reprises d'un même objet aboutit généralement à de petites différences. Par exemple, si vous mesurez plusieurs fois les dimensions d'une pièce avec un mètre ruban, l'hypothèse la plus probable n'est pas de retrouver exactement les mêmes dimensions mais d'observer de légères variations entre les différentes mesures. Plus un instrument de mesure sera précis (par

exemple s'il présente de nombreuses graduations) plus ces variations seront faibles. Ce qui explique ces phénomènes de variation est l'**erreur de mesure**.

Il convient ici de préciser que nous nous situons dans la théorie classique des tests et du **modèle du « score vrai »**. C'est dans ce cadre théorique que se situent la plupart des tests proposés actuellement en France.

Modèle du score « vrai »

La théorie classique des tests différencie le « score vrai » et le « score observé ». Ce que l'on mesure par un test est un score observé. Ce score observé n'est qu'une estimation du score vrai du sujet. Le score vrai est inconnu. La variation entre score vrai et score observé correspond à l'erreur de mesure (voir formule F1)

score observé = score vrai + erreur de mesure (F1)

L'erreur de mesure est aléatoire. Elle se distribue donc selon une courbe de Gauss (loi normale). De même, le score observé se distribue normalement autour du score vrai. Autrement dit, s'il était possible de faire passer un même test à un même sujet un très grand nombre de fois, et qu'il n'y ait aucun effet d'apprentissage, la meilleure estimation du score vrai du sujet serait alors la moyenne des différents scores observés.

Les sources principales de l'erreur de mesure sont les suivantes :

- Effets *propres à l'individu* : état de santé, implication dans le test, réponses au hasard... ;
- Effets *propres au psychologue* : non respect des conditions de standardisation, initiatives malheureuses, erreurs de calcul de scores... On retrouve ci l'importance du respect des conditions de standardisation qui a été soulignée dans la partie précédente ;
- Effets *éventuels du contexte ou de la situation de passation* (caractéristiques de la pièce, bruits éventuels du voisinage...) qui peuvent être plus ou moins propices à la concentration du sujet...

On notera que l'erreur de mesure peut jouer dans les deux sens. Par exemple, si le sujet est un peu fiévreux le jour de passation ou préoccupé par un événement personnel, il est probable alors que son « score observé » sera inférieur à son « score vrai » ; par contre, un sujet qui donne plusieurs réponses au hasard dans un test Q.C.M peut, avec un peu de chance, gagner quelques points et obtenir au final un « score observé » plus élevé que son « score vrai ».

Ainsi le psychologue doit toujours garder à l'esprit que ce qu'il mesure, n'est pas le score vrai du sujet mais n'est qu'une des estimations de celui-ci. Autrement dit il faut toujours considérer que le sujet pourrait avoir un score vrai plus faible ou plus élevé que le score effectivement observé. Il est possible d'estimer cette marge de variation par le calcul d'un intervalle de confiance. Il est en effet possible d'estimer l'erreur de mesure d'un test à partir de son coefficient de fidélité et de calculer alors les limites d'un intervalle dans lequel se trouvera le score vrai. Cette erreur de mesure doit être fournie par les concepteurs du test et figure le plus souvent dans les manuels sous le terme « S.E.M » (Standard Error of Measurement) ou « E.M ».

En fonction du seuil de confiance choisie, le psychologue peut ainsi calculer différents intervalles de confiance grâce aux formules suivantes F2 et F3.

Au seuil de .10 (qui correspond à 10 chances sur 100 de se tromper, c'est-à-dire que sur 100 mesures effectuées sur le même sujet, 90 se situeront dans cet intervalle) :

score vrai = score observé +/– 1,65 x EM (F2)

Au seuil de .05 (qui correspond à 5 chances sur 100 de se tromper : sur 100 mesures, 95 se situeront dans cet intervalle) :

score vrai = score observé +/– 1,96 x EM (F3)

Par exemple, pour l'échelle de Wechsler pour enfants (WISC-III), le manuel français indique l'erreur standard de mesure pour chaque subtest et pour les indicateurs de QI. Par exemple, sur l'ensemble des groupes d'âges, cette erreur de mesure est de 3,54 points pour le QI Total, de 3,85 points pour le QI Verbal et de 5,02 points pour le QI Performance. À partir de ces estimations de l'erreur de mesure il est possible de calculer, pour chaque QI observé un intervalle de confiance. Par exemple, pour un QI Total observé de 105, il y a donc 95 chances sur 100 (seuil de .05) pour que le « score vrai » du sujet se situe entre [105 – (1,96 x 3,54)] et [105 + (1,96 x 3,54)], c'est-à-dire entre 98 et 112.

Si l'on pense que le seuil est trop exigeant et que l'on choisisse alors un seuil de .10, on va alors réduire cet intervalle. Le score vrai se situant alors entre 99[1] et 111[2].

Au seuil .10 nous observons bien que l'intervalle est un peu plus réduit, ce qui s'explique par le fait que l'on prend alors un risque plus important qu'au seuil de .05.

Cet exemple nous indique qu'il est nécessaire de moduler le niveau de précision de la mesure effectuée, c'est-à-dire le score observé, et qu'il est préférable, et plus valide, de caractériser le niveau de performance du sujet par un intervalle de confiance que par un score précis.

De plus, un score unique présente l'inconvénient de créer artificiellement des différences entre des sujets. Ainsi, Huteau et Lautrey indiquent que :

> « On voit combien sont contestables les pratiques qui préconisent des traitements différents pour des individus que ne séparent que quelques points de Q.I. » (Huteau et Lautrey, 1999a, p. 105).

En effet, comment être sûr qu'un QI observé de 81 sur un sujet A reflète réellement des capacités intellectuelles supérieures à celles observées sur un sujet B qui obtiendrait un QI de 79 ?

Même s'il est possible, et souhaitable, de calculer un intervalle de confiance pour tout score observé, très peu de tests facilitent et/ou encouragent ce calcul.

À notre connaissance ce calcul n'est proposé explicitement que dans les échelles de Wechsler qui incitent fortement le psychologue à calculer cet intervalle et à le faire figurer en première page du protocole du sujet.

Dans les autres tests l'erreur type de mesure est indiquée dans le manuel mais ensuite, le plus souvent, les auteurs n'y font plus jamais référence ! Pourtant, quand il s'agit de situer le sujet dans un étalonnage, il nous semble essentiel de prendre en compte cette marge d'erreur.

Peut-on améliorer la fidélité d'un test ?

On peut retenir le principe général suivant : plus une épreuve comporte d'items, plus elle sera représentative de la dimension à évaluer, donc plus sa fidélité augmente. En effet, la consistance de la mesure s'améliore avec le nombre d'items. Mais les concepteurs de tests sont limités ici par des

1. [105 – (1,65 x 3,54)]
2. [105 + (1,65 x 3,54)]

considérations pratiques relatives au temps de passation : plus une épreuve comporte d'items et plus la durée de passation est élevée, et l'on sait qu'une épreuve trop longue a peu de chances d'être utilisée par les praticiens. Il convient alors, dans la phase d'élaboration d'un test de considérer à la fois les contraintes de fidélité et les contraintes pratiques de durée de passation. On notera à ce sujet, et c'est un paradoxe, que de nombreux éditeurs de tests, notamment sur le marché des tests en ligne, trouvent un argument commercial dans la brièveté (parfois extrême) des passations de leurs épreuves. Cet argument doit pourtant alerter l'utilisateur du risque d'affaiblissement de la validité que cela entraîne.

Les différentes formes de fidélité

On distingue principalement trois formes de fidélité : la fidélité dans le temps, la fidélité interne et la fidélité de la cotation.

1. *La fidélité dans le temps* (ou stabilité temporelle)
   Un test doit donner des résultats équivalents quel que soit le moment de passation, à condition bien entendu de prendre en compte les éventuels effets d'apprentissage entre les passations. Cette fidélité peut se mesurer par deux applications successives du même test aux mêmes sujets : méthode dite du test/retest. Un test sera fidèle s'il indique les mêmes résultats pour chaque sujet, au plutôt le même classement des sujets, dans les différentes mesures effectuées avec ce même test. L'indicateur de cette fidélité est ici le coefficient r de B-P, appelé également dans cette situation coefficient de stabilité ou de constance. Pour évaluer la fidélité d'un test par la méthode test/retest il est fortement conseillé de ne pas dépasser un intervalle de six mois entre les passations, car dans le cas d'un intervalle trop long des variables parasites peuvent intervenir et biaiser le calcul des indicateurs. Généralement les tests d'intelligence présentent une bonne fidélité temporelle avec des coefficients de stabilité autour de .90 (Huteau et Lautrey, 1999a, p. 101).
2. *La fidélité interne* (ou homogénéité interne)
   Il s'agit ici de vérifier que tous les items d'une même épreuve mesurent bien la même dimension. Par exemple, si un test souhaite évaluer le raisonnement déductif, tous les items de ce test doivent faire appel à ce type de raisonnement. Mais les items sont différents les uns des autres (par leur contenu, par le problème à résoudre...) et il faut alors vérifier que, malgré ces différences, tous ces items évaluent bien la même

dimension, la même variable psychologique. Si ce n'est pas le cas, certains items mesurent, au moins en partie, autre chose que ce que mesurent les autres items et l'homogénéité de l'épreuve n'est alors plus garantie.
Pour vérifier cette fidélité interne (ou homogénéité interne) on distingue principalement deux méthodes[1] :

– La *première méthode*, diviser l'épreuve en deux parties ou méthode *split-half* (partage par moitié). Par exemple, en réunissant les items pairs d'un côté et les items impairs de l'autre, on aboutit à deux formes parallèles de l'épreuve, mais réduite chacune à 50 % des items de l'épreuve originale. L'indicateur de ce type de fidélité est encore un coefficient de corrélation, un r de B-P, appelé ici coefficient d'homogénéité. Attention ici dans l'interprétation de la valeur du r car comme chaque partie ne comporte que la moitié des items de l'épreuve complète et que, comme nous l'avons déjà signalé, la fidélité dépend (en partie) du nombre d'items, la valeur du coefficient d'homogénéité peut en être affectée. De plus, cette méthode présente l'inconvénient de ne prendre en compte qu'un seul type de répartition des items (items pairs/impairs dans notre exemple) alors que de nombreuses autres partitions des items sont possibles. C'est pour cette raison qu'il est préférable d'utiliser la seconde méthode ;
– La *seconde méthode* présente l'avantage de prendre en compte l'ensemble des répartitions possibles des items de l'épreuve en deux parties égales. En fonction du type d'item il est possible d'utiliser le coefficient dit « KR20 » de Kuder-Richardson (pour des items dichotomiques) ou le coefficient alpha de Cronbach. On considère généralement que l'homogénéité interne est satisfaisante si l'indicateur est ici au moins égal à .80 et on peut considérer l'homogénéité comme acceptable si la valeur de l'indicateur est supérieure à .70 (d'après Vrignaud, 2002b ; voir également Rolland, 2001). En dessous de cette valeur on peut considérer l'homogénéité comme trop faible. Mais attention, il faut encore moduler ces critères en fonction du nombre d'items : toutes choses égales par ailleurs, l'alpha de Cronbach est lui aussi dépendant du nombre d'items (à homogénéité équivalente, un test A possédant plus d'items qu'un test B présentera une valeur plus élevée de cet indicateur alpha).

1. En plus de ces deux méthodes principales il existe d'autres possibilités de vérifier l'homogénéité comme par exemple les techniques d'analyses factorielles (voir Dickes *et al.*, 1994).

3. *La fidélité de la cotation*
   Il s'agit ici de la troisième forme de fidélité que doit présenter un test psychologique. Cette fidélité inter-juges est requise car, pour que la mesure soit fiable, il faut que face à un même protocole (à une même *performance* du sujet) des psychologues différents aboutissent au même résultat, c'est-à-dire au même score. Cette exigence peut sembler évidente et allant de soi mais c'est loin d'être le cas. Nous prendrons deux exemples :

   – Le *premier exemple*, bien connu des téléspectateurs, est celui de l'évaluation des épreuves sportives de patinage artistique. Chacun d'entre nous a pu, au moins une fois, être témoin des écarts de notation entre des juges face à une même performance. Rappelons que dans ces compétitions l'évaluation de chaque juge est publique et consiste à brandir une note à la fin de la prestation de chaque sportif. Et le cas le plus rare est bien celui où tous les juges indiquent la même note ! On observe le plus souvent des écarts de notation entre les juges, alors qu'ils ont pourtant été témoins de la même performance du candidat. Même lorsque ces écarts sont minimes, ils existent ;
   – Le *second exemple* fait référence à un domaine moins connu qui est celui des évaluations scolaires. Les expériences de multi-correction (une même copie, ou un même lot de copie, corrigée par différents enseignants) sont rares. Mais quand elles sont réalisées, elles aboutissent à la mise en évidence de variabilité dans la notation. En effet, tous les travaux de docimologie et cela depuis fort longtemps, montrent, une faiblesse de la fidélité inter-juges dans les évaluations scolaires (voir sur ces points Piéron, 1963, Noizet et Caverni, 1978, et Chartier, 2005).

Les évaluations psychologiques, et plus particulièrement les tests d'intelligence, ne doivent pas présenter ce type de biais. Comme nous l'avons déjà indiqué, du fait même de leur conception, ils garantissent cette forme de fidélité. En effet, dans le cas de Q.C.M, la cotation est simple. Elle est même quelque fois « automatisée » (feuille auto-corrective ou cotation informatisée), ce qui réduit très fortement la possibilité de biais. Dans le cas de réponse à construire, le psychologue doit trouver dans le manuel du test des indications précises afin de pouvoir procéder à la cotation du protocole du sujet avec confiance. Par exemple, le manuel doit indiquer des exemples de bonnes et de mauvaises réponses afin d'éviter toute ambiguïté dans la cotation. Un bon exemple ici concerne les échelles de Wechsler. En

effet, dans les échelles verbales de ces épreuves, certains subtests prennent la forme de réponses à construire avec une cotation précise et assez fine car elle aboutit à des scores de 0, 1 ou 2 points. Pour procéder à cette cotation, le psychologue doit se référer au manuel qui propose, d'une part, les règles générales de définition des trois catégories de réponse, d'autre part, pour chaque item les réponses les plus fréquentes (observées lors de la phase d'expérimentation) avec les cotations correspondantes. Par exemple, pour le subtest vocabulaire du WISC-III, on accorde 0, 1 ou 2 points selon le type de réponse :

– 0 point : réponse incorrecte,
– 1 point : réponse correcte mais non généralisable,
– 2 points : réponse correcte et généralisable.

Et le manuel indique, pour chaque item, une liste de réponses possibles avec les scores à attribuer. Le psychologue dispose ainsi de tous les éléments pour procéder avec confiance à la cotation du protocole.

Comme nous l'avons déjà signalé, cette fidélité de la cotation a été évaluée pour la version WISC-III avec le calcul d'une corrélation entre les cotations indépendantes de plusieurs protocoles par deux psychologues avec au final un r de BP quasiment parfait (r = .99).

Comme cet exemple le prouve, la fidélité de la cotation des tests d'intelligence est garantie, même lorsque l'épreuve n'est pas de type Q.C.M, à condition toutefois que le psychologue suive scrupuleusement les indications de correction fournies dans le manuel et que les réponses soient courtes.

Face à une épreuve présentant des questions ouvertes, le psychologue doit donc s'intéresser de très près aux consignes de correction et aux études présentées dans le manuel qui doivent vérifier ce type de fidélité.

3. La notion de sensibilité

L'objectif principal d'un test est bien d'observer des différences interindividuelles. Cette capacité de différenciation des sujets correspond à la notion de sensibilité. La sensibilité représente le pouvoir discriminatif de l'instrument de mesure. Un test est sensible s'il permet bien de distinguer des sujets de niveaux différents sur une même dimension psychologique, comme par exemple l'intelligence. Un des postulats de base en psychométrie, et plus globalement en psychologie, et valable quel que soit le type de test,

consiste à considérer que les dimensions psychologiques se répartissent dans la population selon une loi normale (une courbe de Gauss) comme l'indique la figure 2.1.

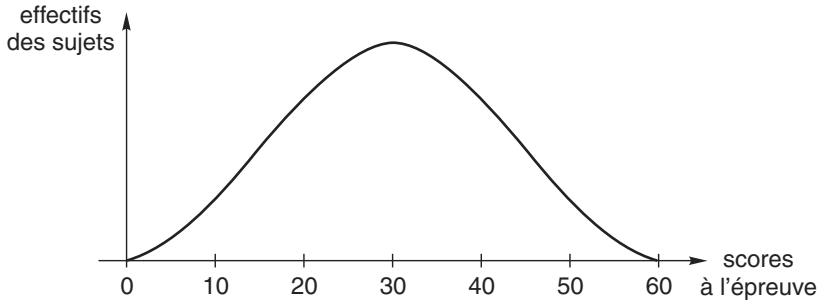

Figure 2.1
Exemple d'une répartition de scores conformes à une courbe de Gauss (D'après Guéguen, 2005, Statistiques pour psychologues, Dunod, p. 80).

Soit une épreuve notée sur 60 points, la répartition théorique des scores des sujets est présente sur la figure 2.1.Un test doit donc aboutir à une telle répartition des sujets : une faible proportion de sujet doit se trouver sur la gauche de la courbe (c'est-à-dire obtenir des scores faibles au test), symétriquement une proportion identique doit se situer sur la droite (scores élevés), avec vers le centre, une majorité de sujets (scores autour de la moyenne), et une décroissance progressive des effectifs des sujets du centre vers les deux extrémités. Dans la phase d'élaboration d'une épreuve, les items sont donc sélectionnés pour assurer cette discrimination entre les sujets. Un test d'intelligence comportera ainsi des items de niveaux de difficulté variables de manière à différencier efficacement les sujets en fonction de leur réussite. Ici va intervenir la notion de difficulté de l'épreuve. Pour assurer une sensibilité maximale, un test doit présenter un niveau moyen de difficulté par rapport au niveau de la population auquel il est destiné. Il faut éviter par exemple « l'effet plafond », qui correspond à une épreuve trop facile (tous les sujets obtiennent alors de bons scores, il n'y a donc pas discrimination), ainsi que l'effet inverse, « l'effet plancher », qui correspond à une épreuve trop difficile dans laquelle tous les sujets obtiennent des notes basses. Dans ces deux situations la différenciation des sujets n'est pas réalisée de façon satisfaisante et le test n'a alors que peu d'utilité.

Cette finesse de la mesure est également liée au nombre d'items de l'épreuve : plus une épreuve comportera d'items, plus elle permettra d'effectuer une différenciation fine entre les sujets.

Enfin, cette sensibilité est liée au pouvoir discriminant des items (voir plus loin).

Dans les tests d'intelligence, on accorde le plus souvent un point par item réussi et on additionne ces points pour obtenir le score brut. Les sujets doivent donc se différencier sur ce score.

4. La notion de validité

Principes

Cette notion est fondamentale.

> « En psychométrie, la validité a toujours été considérée comme le concept le plus fondamental et le plus important. » (Angoff, 1988, cité par Laveault et Grégoire, 1997, p. 189.)

Qu'est-ce que la validité ? Elle correspond à ce que mesure le test. Par exemple, une balance mesure bien un poids (ou une masse) et non un volume. De même un test de raisonnement particulier doit mesurer ce type de raisonnement et seulement ce type de raisonnement.

Mais cette validité ne va jamais de soi, elle doit toujours être démontrée, vérifiée. Des éléments de validation du test doivent être présentés par ses concepteurs. L'utilisateur du test doit pouvoir vérifier dans le manuel la présence et la pertinence de ces éléments de validation. Il s'agit le plus souvent de résultats de recherches menées lors de la phase d'expérimentation de l'épreuve. Mais ces recherches sont souvent en nombre assez réduit lors de la première édition d'un test et il convient alors, afin de compléter ces premières preuves de validité, de prendre en compte les informations ultérieures sur le test (études, recherches, articles...) publiées après la publication du manuel du test (et qui ne figurent donc pas dans ce manuel). Un bon exemple est celui des matrices de Raven : depuis l'élaboration en 1938 de la première version de ces matrices, des études viennent régulièrement s'ajouter aux éléments de validation déjà connus (voir sur ce point Raven, 2001, qui présente une synthèse des nombreuses recherches sur ce test).

Ainsi, progressivement, les connaissances sur ce que mesure une épreuve, et éventuellement aussi sur les aspects qu'elle ne mesure pas, vont se cumuler et enrichir notre connaissance du test. C'est pour cette raison que certains

auteurs préfèrent actuellement utiliser la notion de *validation*, qui exprime ce processus cumulatif :

> « Dans la conception actuelle, le terme de validité est abandonné au profit de celui de validation. Sous ce changement terminologique qui peut sembler anodin, on trouve en fait un changement radical de conception : la validation devient un processus de recherche continu qui s'appuie sur un faisceau convergent d'arguments et de preuves. » (Dickes *et al.*, 1994, p. 49).

Tout psychologue devrait ainsi se tenir informé des résultats des recherches sur les outils qu'il utilise afin de mettre à jour ses connaissances (voir en Annexes le code de déontologie et les recommandations internationales).

Différents types de validité

On distingue habituellement trois grandes formes de validité : la validité de contenu, la validité critérielle et la validité théorique.

➤ La validité relative au contenu du test

Dans cette première forme de validité, il s'agit de vérifier dans quelle mesure le test est représentatif du domaine à évaluer. Une définition préalable du domaine est nécessaire et doit comporter des informations précises sur les aspects, ou facettes, censés être évalués par le test. Pour vérifier cette forme de validité une analyse de la liaison entre le domaine, ou les sous-domaines, visés par le test et le contenu du test doit être menée (format et contenu des items, type de réponse...).

Pour garantir un bon niveau de validité de contenu, un test doit comporter un échantillon représentatif des tâches caractérisant le domaine considéré. Cette forme de validité est particulièrement pertinente pour les évaluations de connaissances. Par exemple, un test de mathématiques pour des élèves de niveau de la classe de 3e de collège aura une bonne validité de contenu si les exercices (items) qu'il contient correspondent à un échantillon représentatif du programme de mathématique de ce niveau scolaire. Elle est en revanche moins pertinente dans le domaine des tests d'intelligence car il y est plus difficile de sélectionner un tel échantillon représentatif de *l'univers des items*, c'est-à-dire de l'ensemble des items constituant le domaine. Comment s'assurer de la représentativité de l'échantillon des items qui constituent le test si l'on ne connaît pas l'ensemble des items possibles ? On peut noter ici les travaux originaux de Dickes sur la définition de l'univers des items

pour la tâche des cubes de Kohs (voir sur ce point Dickes, 1988 et Dickes, Houssemand et Reuter, 1996) mais ce type de recherche, et nous pouvons le regretter, reste une exception. De ce fait, concernant les tests d'intelligence, on accordera plus d'importance à la validité théorique (voir plus loin).

Il faut donc toujours garder à l'esprit que le test, et les items qui le composent, ne sont qu'un échantillon des situations caractéristiques du domaine considéré et que la représentativité de ces items n'est généralement pas évaluée de façon précise. Le plus souvent, à partir de références théoriques (voir plus loin la notion de validité théorique), le concepteur du test va sélectionner un certain type de tâche (donc un certain type d'items) qui sera en rapport direct avec ce cadre théorique. Mais dans cette sélection d'item, la représentativité est plus ou moins bien assurée. C'est ce qui explique, par exemple, qu'un test d'intelligence présente toujours une spécificité (on peut faire ici le lien avec le facteur spécifique distingué par Spearman, voir chapitre 1 de ce livre) et que, même à l'intérieur d'un cadre théorique identique, une épreuve ne sera jamais parfaitement équivalente à une autre : chacune ayant sélectionné, parmi l'ensemble des possibles, certaines situations qui vont alors définir les caractéristiques des items du test.

Ainsi par exemple, si l'on prend deux tests de facteur g, le D2000 et le SPM de Raven, et bien que leur cadre théorique soit identique (référence à l'approche de Spearman, avec comme objectif commun d'évaluer le facteur g) le type d'items est différent (domino pour le D2000, matrice pour le PMS), ainsi que les modalités de réponse (réponse à construire pour le D2000, réponse à choisir QCM pour le SPM). Ces deux épreuves évaluent bien la même dimension (ici, le facteur g) mais à partir de situations différentes. Cette même dimension peut en outre être également évaluée, et de manière aussi satisfaisante, à l'aide d'autres tests. Autrement dit, chaque épreuve de facteur g propose des tâches qui ne sont qu'un échantillon de l'ensemble des tâches permettant d'évaluer le facteur g. Ce principe est valable quel que soit le cadre théorique de référence.

Nous avons pris ici pour exemple deux tests de facteur g dont la validité est éprouvée et reconnue. Mais il faut cependant être toujours attentif au contenu d'un test, à sa validité de contenu, car elle va, tout au moins en partie, conditionner la généralisation que l'on peut faire à partir des résultats obtenus dans le test. Ainsi, nous verrons plus loin que certains biais d'évaluation sont en rapport direct avec le contenu des items.

➤ La validité critérielle en référence à un critère externe

Il s'agit ici d'analyser les liaisons existant entre le niveau de réussite au test et le niveau de réussite dans une autre situation prise comme critère. Ce peut-être par exemple le lien entre les résultats à un test d'intelligence et la réussite scolaire, ou avec l'obtention d'un examen, ou encore la liaison avec l'adaptation à un poste de travail... Un test a une bonne validité critérielle lorsqu'il présente une liaison élevée avec le critère considéré.

En fonction de l'intervalle de temps entre les deux mesures, on distingue la validité *concomitante* et la validité *prédictive.*

– *La validité concomitante* (ou concurrente[1]) rend compte de la liaison entre le test et le critère lorsque les deux mesures se situent dans un même temps. Il va s'agir, par exemple, de procéder à la passation d'un test d'intelligence au 1er trimestre scolaire et d'analyser les liaisons avec les résultats scolaires du 1er trimestre ;
– *La validité prédictive* consiste à évaluer les sujets, dans un premier temps avec le test puis, après un intervalle plus ou moins long, de recueillir, dans un second temps, les données sur le critère. On cherche ainsi à savoir si le test permet de prédire, avec plus ou moins de confiance, les résultats obtenus sur le critère. Pour reprendre notre exemple il s'agira alors d'analyser, par exemple, les liaisons entre le test passé au 1er trimestre et l'obtention d'un diplôme en fin de 3e trimestre : les élèves qui avaient eu de bons résultats au test ont-ils mieux réussi le diplôme que les élèves qui avaient eu de plus faibles résultats au test ? Si c'est le cas, alors ces résultats iront dans le sens d'un bon niveau de validité prédictive du test car il sera possible de prédire l'obtention du diplôme, et plus généralement de prédire le niveau dans le critère, à partir des résultats au test. Le test peut alors faire office de prédicteur. L'indicateur de validité critérielle est le plus souvent un coefficient de corrélation.

Deux points sont à retenir ici : d'une part, plus cette validité est élevée et plus la prédiction sera précise, d'autre part, la qualité de cette prédiction réalisée à partir d'une seule variable (le test) peut être améliorée en prenant en compte plusieurs variables au lieu d'une seule (prendre en compte, par exemple, les résultats à plusieurs tests pour prédire un évènement).

Pour les tests d'intelligence, cette validité est, en moyenne de .50 avec des indicateurs de réussite scolaire (Reuchlin, 1991) ainsi qu'avec des

1. On trouve également le terme de concourrente.

indicateurs de réussite professionnelle (Robertson & Smith, 2001 ; Salgado, 1999 ; Smith & Hunter, 1998).

Ces valeurs de validité pronostique dépendent à la fois du test (toute chose égale par ailleurs, deux tests peuvent présenter des valeurs différentes de validité pronostique par rapport à un même événement) mais dépendent également du type d'indicateur utilisé pour le critère. Par exemple, dans le cas de la validité prédictive des tests d'intelligence par rapport à la réussite scolaire, les validités (c'est-à-dire les coefficients r de B-P) sont généralement plus élevées quand on utilise, pour le critère de réussite scolaire, des épreuves normalisées de connaissances que lorsqu'on utilise les notes quotidiennes des enseignants (en raison, principalement, d'une fidélité plus faible de ces notes).

➤ La validité théorique en référence à un concept ou à un modèle théorique

On parle aussi de validité de *construct*, de validité conceptuelle, de validité hypothético-déductive.

Ce type de validité questionne directement les présupposés théoriques qui sont à la base de l'épreuve. Effectivement toute épreuve est basée sur des idées, sur des concepts qui peuvent être plus ou moins élaborés. Ces idées sous jacentes, ces théories, conditionnent la validité d'un test :

> « Les tests valent ce que valent les idées qui ont présidé à leur construction. » (Huteau & Lautrey, 1997, p. 3).

C'est cette forme de validité qui permet de donner du sens à ce qui a été évalué, de donner une signification à un score.

On parle de validité *convergente* et de validité *divergente* : un test valide doit présenter une corrélation forte avec une épreuve qui mesure le même domaine (validité convergente), et une corrélation nulle, ou faible, avec une épreuve évaluant un autre domaine ou une autre dimension indépendante (validité divergente). Par exemple, deux tests d'intelligence doivent présenter une corrélation élevée car ils sont censés évaluer tous les deux une même dimension : l'intelligence (validité convergente). Par contre, en l'absence d'hypothèse spécifique à ce niveau, ils ne doivent pas présenter un tel niveau de liaison avec, par exemple, un test de personnalité, car il s'agit d'un domaine différent, relativement indépendant de l'intelligence (validité divergente).

Un test doit ainsi spécifier les bases théoriques sur lesquelles il repose et présenter des informations qui viennent étayer les propos. Il s'agit le

plus souvent de situer le test parmi les modèles théoriques de référence (voir le chapitre 1) et de confronter les résultats d'un groupe de sujets à des tests comparables. Par exemple : un test censé évaluer le facteur général d'intelligence (facteur G) doit présenter une forte corrélation avec un test déjà connu (et validé) qui évalue ce même facteur (validité convergente). Si ce n'est pas le cas, si les liaisons ne sont pas assez fortes entre les deux épreuves, alors ce nouveau test ne peut pas affirmer qu'il évalue lui aussi le facteur g : sa validité théorique n'est pas assurée.

Comme nous venons de le voir dans l'exemple, on retrouve ici comme indicateur de validité le coefficient de corrélation. Mais attention dans l'interprétation de ces coefficients : on ne pourra jamais obtenir ici des valeurs très proches de 1 car il existe une relation entre fidélité et validité : les fidélités réciproques de deux mesures déterminent les limites supérieures de leur corrélation. Autrement dit, la corrélation maximale entre deux tests est limitée par le fait qu'aucun d'eux n'est une mesure parfaitement fidèle (voir l'exemple de Laveault et Grégoire, 1997, p. 205).

Une autre facette de la validité théorique est la validité *structurale* d'un test. De quoi s'agit-il ? Dans le cas où le modèle théorique de référence envisage plusieurs dimensions, comme par exemple dans les tests de Wechsler qui distinguent échelle verbale (et QIV) et échelle de performance (et QIP), on doit retrouver des indicateurs statistiques qui viennent confirmer (valider) cette distinction théorique. Plusieurs méthodologies sont utilisables, et principalement les techniques d'analyse factorielle.

Partons d'un exemple : l'épreuve de Wechsler WISC-III est organisée autour de deux échelles afin de pouvoir calculer ces deux Q.I. Pour valider cette structure, les subtests d'une échelle doivent alors présenter entre eux des corrélations plus élevées qu'avec les subtests de l'autre échelle. Ce qui doit se traduire dans une analyse factorielle par la mise en évidence de deux facteurs distincts, qui viennent alors confirmer, et valider, le calcul de ces deux indices (QI V et QI P). Autrement dit, on doit retrouver au niveau même de l'organisation des données, le regroupement des subtests en deux blocs, correspondant à la distinction théorique proposée par Wechsler. Effectivement, comme nous le présenterons dans un autre chapitre, nous retrouvons dans le manuel du WISC-III des analyses factorielles qui valident la distinction proposée par l'auteur de ce test. Il s'agit bien ici de la validité *structurale* du test, validité relative à la structure interne de l'instrument.

Un autre exemple de validation de la structure d'un test nous est donné par le test K-ABC (Kaufman et Kaufmann, 1993). Le cadre de référence théorique principal de cette épreuve repose sur des travaux de

psychologie cognitive et de neuropsychologie, menés dans les années 1980, qui proposaient de distinguer deux grands types de processus cognitifs : les processus simultanés et les processus séquentiels.

Les processus simultanés sont utilisés lorsque les caractéristiques de la situation nécessitent de traiter en même temps plusieurs informations, les processus séquentiels correspondent eux à un traitement pas à pas des informations.

Les concepteurs du test K-ABC ont donc souhaité élaborer une épreuve qui permette d'évaluer chaque type de processus. Ils ont alors sélectionné des items pour évaluer les processus simultanés et d'autres items pour évaluer les processus séquentiels. Au final, le K-ABC se présente un peu comme la structure des échelles de Wechsler avec deux échelles[1] distinctes : l'une pour les processus simultanés, l'autre les processus séquentiels.

Mais dans la phase d'expérimentation de leur épreuve, les analyses ont montré qu'un des subtests, conçu à l'origine pour faire partie de l'échelle des processus séquentiels, était en fait corrélé plus fortement avec les items de l'autre échelle. Pour conserver un bon niveau de validité structurale à leur épreuve les auteurs ont alors déplacé ce subtest vers l'échelle de processus simultanés (Kaufman *et al.*, 1993, p. 55). S'ils n'avaient pas modifié ainsi la structure de leur épreuve expérimentale, la validité structurale du test en aurait été affectée.

Cette validité structurale est également exigée dans tous les domaines concernés par les tests, comme par exemple dans les questionnaires d'intérêts basés sur la théorie de Holland, où il va s'agir alors de retrouver les six types RIASEC postulés par ce modèle théorique, organisés en hexagone (Vrignaud et Bernaud, 2005).

Les tests d'intelligence doivent donc présenter des éléments de validation selon ces trois axes :

– validité de contenu,
– validité critérielle,
– validité théorique.

Dans la réalité, comme nous l'avons indiqué, les manuels de tests accordent une part plus importante aux deux dernières formes de validité. Mais, comme nous le verrons dans la partie consacrée à la présentation de tests, ces informations sont plus ou moins complètes selon les épreuves.

1. En réalité le K-ABC comporte également une échelle de connaissance mais nous n'en parlerons pas ici afin de faciliter la compréhension de l'exemple.

Pour terminer, signalons qu'une autre forme de validité est souvent citée : la validité apparente (*face validity*). Cette validité est en quelque sorte une validité « de surface » (on parle aussi de validité écologique) et correspond à une analyse *intuitive* du test. C'est par exemple, une analyse du contenu du test effectuée par un juge non spécialiste du domaine qui aboutirait à un jugement de type « *ce test évalue l'intelligence parce que ça se voit !* ». Ce type de validité, même s'il est pertinent dans le cadre par exemple de la phase de restitution des résultats, n'est bien entendu pas suffisant. Un test qui ne présenterait que ce type de validité ne serait pas automatiquement valide, car il n'y a ici aucune vérification objective sur ce qui est réellement mesuré par ce test.

5. L'analyse des items

Dans le processus de création d'un test, cinq étapes peuvent être distinguées (Laveault et Grégoire, 1997) :

- la détermination des utilisations prévues du test,
- la définition de ce que l'on souhaite mesurer,
- la création des items,
- l'évaluation des items,
- la détermination des propriétés métriques du test définitif.

La forme expérimentale d'un test doit comporter un nombre suffisant d'items de manière à pouvoir sélectionner les items les plus pertinents qui constitueront la version définitive du test. Par exemple, si un test doit comporter au final 30 items, il faudra en créer environ 45, mener une expérimentation tous et ne retenir que les 30 meilleurs.

Sur quelles bases est effectuée cette sélection d'items ? Après avoir analysé la fidélité de la mesure, et éventuellement retiré les items qui ont été identifiés comme responsables d'une éventuelle faiblesse de l'épreuve à ce niveau, chaque item va être caractérisé par deux indicateurs principaux : un indice de difficulté et un indice de discrimination. C'est souvent, en grande partie, à partir de ces deux indicateurs que la sélection finale des items sera réalisée. Ce qu'il faut noter c'est que, pour chaque item constituant l'épreuve définitive, les valeurs de ces indicateurs doivent figurer dans le manuel.

Il est donc important de connaître ce que représentent ces indicateurs. De plus, dans certains cas, il peut être nécessaire de revenir vers ces indices pour mieux comprendre la réussite ou l'échec d'un sujet à un item particulier.

L'indice de difficulté

Cet indicateur est très simple à calculer et à interpréter : à partir du nombre de sujets ayant réussi l'item, et le nombre total de sujets, on peut calculer une fréquence de réussite à l'item. Cette fréquence de réussite, exprimée le plus souvent sous la forme d'un pourcentage, représente l'indice de difficulté de l'item.

> ***Exemple***
>
> Si 56 sujets sur 109 réussissent un item A, l'indicateur de difficulté de A est donc de 56/109 = 0,514 soit 51,4 % de réussite. Autrement dit c'est un item de difficulté moyenne car il a été réussi par un peu plus de la moitié des sujets de cet échantillon. Et si seulement 32 sujets réussissent l'item B, l'indice de difficulté de B est donc de 32/109 = 0,294 soit 29,4 %. L'item B, qui présente un indicateur de réussite plus faible, est donc plus difficile que l'item A.

Cet indicateur est en lien direct avec la notion de sensibilité : il permet de juger de la capacité de l'item à différencier les individus. L'indice de difficulté est directement dépendant du niveau de l'échantillon. On considère qu'un item a un bon pouvoir différenciateur lorsque l'indice de difficulté est proche de 50 % (cas de l'item A de notre exemple). Une valeur plus élevée témoigne d'un niveau de difficulté plus faible et inversement, une valeur plus faible témoigne d'un niveau de difficulté plus élevé (cas de l'item B de notre exemple).

Pour assurer un bon niveau de sensibilité du test on considère qu'il faut que l'épreuve, au total, présente un niveau de difficulté moyen. Pour arriver à ce résultat on sélectionne les items dont la fréquence de réussite est comprise entre 30 et 70 % environ. Mais l'épreuve doit comporter également des items plus faciles, qui seront placés généralement en début d'épreuve (afin de motiver les sujets), et qui permettent de distinguer les sujets de faible niveau, et des items plus difficiles, placés généralement plutôt en fin d'épreuve, qui serviront à différencier les sujets de niveau de compétence plus élevée.

Cet indice de difficulté, au niveau de l'item comme au niveau global de l'épreuve, dépend donc des caractéristiques du groupe de sujets sur lequel s'effectue la passation : le niveau de difficulté d'un item, ou d'un test, peut ainsi varier en fonction du niveau des sujets de l'échantillon. C'est ce qui

explique que, pour un même test, il existe souvent plusieurs étalonnages : chacun correspondant à un groupe précis de sujets (voir plus loin la notion d'étalonnage).

Cette dépendance entre caractéristiques des sujets et caractéristiques des items ne permet pas d'établir des échelles de mesures absolues. Pourtant ce type d'échelle présenterait un certain nombre d'avantages : comparaison possible de sujets différents sur une même échelle, comparaison de sujets n'ayant pas passé les mêmes épreuves, calibrage des items pour constituer des banques d'items... C'est en grande partie pour cette raison que d'autres modèles de mesure, comme les Modèles de Réponse à l'Item (ou M.R.I), ont été développés (voir plus loin une présentation de ces modèles).

L'indice de discrimination

La discrimination d'un item représente sa capacité à différencier les sujets qui obtiennent un niveau élevé de réussite à l'ensemble du test des sujets qui présentent un niveau plus faible de réussite. On parle du pouvoir discriminant d'un item. Un bon item est ici un item qui permet bien de distinguer les sujets sur leur niveau de réussite globale à l'épreuve.

On analyse ainsi, au niveau de l'item, la relation entre la réussite à cet item et le score total obtenu au test. On cherche bien entendu une liaison forte : les sujets ayant réussi l'item doivent présenter en moyenne un score total plus élevé que les sujets ayant échoué à l'item.

L'indicateur utilisé ici est la corrélation item/test. Il s'agit plus précisément de la corrélation point bisérial entre l'item (codé en 0/1) et le score total, corrélation corrigée pour la présence de l'item dans le score (cette corrélation revient à calculer la corrélation entre l'item et le score total sans prendre en compte l'item considéré). Un item qui présente une valeur élevée à cet indice sera un item à conserver. Au contraire, un item qui présente une valeur faible, sera à exclure.

Mais quelles sont les valeurs seuils ? Il est assez difficile de trouver des valeurs seuils dans la littérature, car, comme nous l'avons déjà indiqué, d'autres variables sont à prendre en compte. On peut néanmoins considérer que cette discrimination est très satisfaisante si l'indice est supérieur à .40 ; qu'elle est satisfaisante entre .20 et .40 ; qu'elle est faible entre .10 et .20 ; qu'elle est insuffisante en dessous de .10 (Vrignaud, 2002b).

Autrement dit, *théoriquement* un test ne devrait comporter aucun item présentant un indice de discrimination inférieur à .10. Mais *pratiquement*

un item présentant un trop faible niveau de discrimination peut néanmoins être conservé à condition qu'il ait un apport significatif à un autre niveau d'analyse (par exemple, on peut décider de conserver un item en raison de son contenu spécifique...).

6. La notion de biais

Avant de définir et d'illustrer cette notion de biais, il faut, d'une part, rappeler que l'usage des tests a été souvent questionné par la présence, réelle ou supposée, de biais sociaux, ou de biais culturels, dans ces épreuves (voir par exemple Bacher, 1982 ; Huteau et Lautrey, 1999a), d'autre part, les tentatives d'élaboration de tests indépendants de la culture : test *culture free* ou *culture fair*.

Actuellement, s'il semble difficile d'élaborer un test qui soit complètement indépendant d'un contexte culturel, il est par contre possible, et hautement souhaitable, de s'assurer de l'absence de biais. On retrouve d'ailleurs cet objectif dans les *recommandations internationales sur l'utilisation des tests*, qui préconise au psychologue de s'assurer de l'«absence de biais systématiques au détriment de l'un des groupes de sujets auxquels le test sera administré » (section 2.2.2, paragraphe d, page 19).

Des études doivent donc être menées sur ce plan et doivent figurer dans le manuel.

Qu'est-ce qu'un biais ?

> « On dit qu'une mesure est biaisée dès lors qu'elle ne mesure pas, ou qu'imparfaitement, ce qu'elle est censée mesurée. On est en présence d'un biais lorsque la mesure met en évidence des différences entre des groupes de sujets et que ces différences ne peuvent être mises en relation avec la ou les variables mesurées. » (Vrignaud, 2002a, p. 626).

Il est important de distinguer ici clairement ce qui est censé être mesuré par le test (la variable ou dimension) qui doit expliquer les différences de performances observées entre les sujets et/ou entre des groupes de sujets, et l'influence éventuelle d'une autre source de variation (un biais) qui pourrait également expliquer certaines différences observées. Par exemple, si la résolution d'un item de test d'intelligence nécessite de connaître un terme

spécifique (ou technique), la réussite ou l'échec à cet item ne dépendra donc plus uniquement du niveau d'intelligence (variable censée être mesurée par le test) mais également de la connaissance ou non de ce terme (variable ici considérée comme un biais : variable parasite). Autrement dit, à niveau comparable d'intelligence, les sujets connaissant ce terme technique seront avantagés par rapport aux sujets ne connaissant pas ce terme. Dans ce cas, cette situation d'évaluation est donc biaisée. Cet exemple renvoie à la notion de validité de contenu.

D'une façon plus générale, on peut retenir qu'un test est biaisé, ou présente un biais, s'il avantage, ou désavantage, de façon systématique un groupe particulier de sujets.

Différents types de biais

Il est possible de distinguer plusieurs types de biais : les biais de construit, les biais de méthode et les biais d'item (Van de Vijver et Poortinga, 1997).

Nous ne pouvons pas ici détailler l'ensemble de ces biais possibles mais nous en indiquerons uniquement les points essentiels (pour approfondir ce sujet voir Vrignaud, 2002a et 2001 ; Grégoire, 2004).

➤ Les biais de construit, ou biais conceptuels

Ce qui est questionné ici, ce sont non seulement les bases théoriques auxquelles le test se réfère (conception *théorique* de l'intelligence sous jacente) mais, aussi, le sens que prend ce modèle théorique dans le contexte social et culturel particulier dans lequel est élaboré le test et dans lequel sera utilisé le test. Ce sont ces modèles de référence qui vont définir les indices à prendre en compte, les caractéristiques des situations d'évaluation...

La fiabilité de ces modèles à des sujets de culture différente n'est pas nécessairement garantie : la définition d'un comportement intelligent peut varier d'une culture à l'autre. Ainsi par exemple, même si l'on a de bonnes raisons de penser que le modèle de l'intelligence proposé par Carroll (voir chapitre 1) est universel (Grégoire, 2004, p. 90), il est fort probable que certains contextes culturels vont accorder plus d'importance à certaines aptitudes spécifiques qu'à d'autres. L'absence de prise en compte de ces spécificités culturelles peut aboutir à ces biais de construit, d'où la nécessité de procéder à des études rigoureuses lors de l'adaptation de tests. On peut citer ici l'exemple de la structure factorielle du test WISC-III qui comporte

quatre Indices Factoriels dans la version américaine mais n'en comprend plus que trois dans la version française (voir le chapitre 3 qui détaille cet aspect).

➤ Les biais de méthode

Un premier biais de méthode concerne ici des biais liés aux échantillons et principalement, pour ce qui nous concerne, les biais d'étalonnage des sujets. Le groupe de référence, qui constitue l'étalonnage, doit présenter des garanties quant à son niveau de représentativité. Sinon, la comparaison des résultats d'un sujet avec ce groupe de référence ne peut pas être effectuée de manière fiable.

Un deuxième biais de méthode repose sur le matériel utilisé. Les sujets ne doivent pas se différencier sur le degré de familiarité avec le matériel du test. Sinon, un biais est ici possible : les sujets présentant un haut niveau de familiarité avec ce matériel, ou avec le type de tâche, peuvent être avantagés. D'où la nécessité de ne pas diffuser le matériel de test.

Enfin, un troisième type de biais de méthode est relatif aux conditions de passation. Par exemple, la situation même de passation peut sembler étrange à certains sujets, ou à des sujets d'une certaine culture, qui peuvent être étonnés que le psychologue reste en face d'eux sans leur apporter une aide (Reuning et Wortley, 1973, cité par Grégoire, 2004). La compréhension de la situation et donc l'implication du sujet dans cette situation peuvent alors être source de biais.

➤ Les biais d'items

Les deux types de biais précédents, biais de construit et biais de méthode, affectaient l'ensemble du test. Par contre le biais d'item peut jouer uniquement sur un item isolé. On parle de biais d'item lorsque, à niveau de compétence comparable des sujets (estimé par le même score total au test), le niveau de difficulté d'un item particulier va varier en fonction des sujets, en raison de l'effet d'une variable parasite.

C'est l'exemple que nous avons déjà présenté en introduction qui illustrait les effets de la connaissance d'un terme technique (variable parasite) dans un test d'intelligence (variable évaluée). Ainsi, un item sera biaisé en faveur ou en défaveur d'un groupe particulier de sujets. Dans ce cas une variable différente de celle qui est censée être évaluée peut intervenir et favoriser un groupe par rapport à un autre.

Pour repérer les effets éventuels d'une variable parasite, plusieurs méthodes sont utilisables dans le cadre de l'étude du Fonctionnement Différentiel de l'Item (F.D.I) :

– la statistique de Mantel-Haenszel ;
– la régression logistique ;
– l'approche par les Modèles de Réponse à l'Item (M.R.I) ;
– l'approche de Stout.

(Pour une présentation détaillée de ces approches Vrignaud, 2002a. Pour des exemples précis d'analyse de FDI : Vrignaud, 2001).

Quelle que soit la méthode utilisée, l'objectif est le même : repérer les items présentant un F.D.I. Ensuite, plusieurs solutions sont envisageables :

– Retirer les items biaisés ;
– Modifier les caractéristiques de l'item de manière à annuler les effets de la variable parasite ;
– Analyser le test dans son ensemble pour vérifier l'effet cumulé des différents items biaisés.

Par exemple, dans l'expérimentation de la version WISC-III des échelles de Wechsler, une analyse de F.D.I a été réalisée en comparant les résultats d'enfants français et d'enfants belges. Sur le subtest « information » quatre items présentant un F.D.I ont néanmoins été conservés car deux de ces items avantageaient les élèves Français et les deux autres avantageaient les Belges. Dans ces conditions, on peut considérer que les effets cumulés des F.D.I ont tendance à s'annuler (Grégoire, 2000a).

Conclusion sur la notion de biais

Dans cette partie nous avons montré les effets possibles d'un certain nombre de biais potentiels sur la mesure réalisée par un test. Une attention particulière sur ces différents points doit donc être menée dans les différentes phases d'élaboration d'un test. Plusieurs méthodologies existent, plus particulièrement pour identifier les biais d'items, mais on peut remarquer que, en France, ces analyses de biais sont encore assez rares (Vrignaud, 2002a).

7. La notion d'étalonnage

Principes de l'étalonnage

Comme nous l'avons déjà indiqué, l'étalonnage est l'élément qui va permettre de situer les résultats d'un sujet en référence à ceux obtenus par un groupe de sujets comparables (une population de référence). En effet, le score brut du sujet (qui correspond à la somme des points obtenus dans un test) ne veut rien dire en soi. Il ne peut être interprété qu'au regard d'une référence.

Le rôle de l'étalonnage est de fournir cette référence.

Pour étalonner un test il faut le faire passer à un échantillon de sujets de manière à obtenir la répartition des résultats à ce test sur cet échantillon de sujets. C'est cette distribution des résultats qui va servir de référence, de norme. Il faut donc que l'échantillon de sujets présente des caractéristiques comparables à celles de la population de référence (par exemple par rapport au niveau d'étude, au sexe...). Un même test peut être utilisé pour des populations différentes, et généralement on dispose de plusieurs étalonnages pour un même test qui permet de distinguer ces populations (voir notre exemple plus loin).

Bien entendu, comme nous l'avons indiqué lors de la présentation de la notion de sensibilité, les conditions d'application et de cotation doivent être strictement comparables, sinon la comparaison des résultats serait biaisée.

Les caractéristiques des sujets qui constituent les échantillons des étalonnages doivent être soigneusement décrites : nombre de sujets, âges moyens, niveaux scolaires, dates de passation... Le psychologue peut alors sélectionner, parmi les étalonnages disponibles, celui qui lui semble le plus proche des caractéristiques du sujet examiné

Généralement, un étalonnage se présente sous la forme d'un tableau dans lequel figurent les scores bruts et les scores « étalonnés ». Les scores étalonnés prennent souvent la forme de classes et permettent alors de situer le sujet dans l'une des classes. Nous verrons plus loin que les étalonnages peuvent présenter un nombre différent de classes.

Prenons, par exemple, un score de 44 points obtenu dans un test comportant 60 items. Pour ce test nous disposons éventuellement de plusieurs étalonnages en fonction de la profession exercée. Nous avons reproduit dans le tableau 2.1 un étalonnage (fictif) correspondant à une population de niveau cadre.

Tableau 2.1
Étalonnage (fictif) d'un test de 60 items, niveau cadre.

Classes (scores étalonnés)	% théorique de sujets dans chaque classe	Notes Brutes (scores mesurés)
1	4	0-38
2	6,6	39-41
3	12,1	42-44
4	17,5	45-48
5	19,6	49-51
6	17,5	52-53
7	12,1	54-56
8	6,6	57-58
9	4	59-60

Sur ce tableau apparaissent les notes brutes (ou scores bruts) dans la troisième colonne, les classes (ou notes étalonnées) dans la première colonne, et le pourcentage de sujets appartenant à chaque classe dans la colonne centrale (colonne % théorique). Le principe de l'étalonnage est de regrouper certains scores bruts au sein d'un même score étalonné. Par exemple ici, une note *brute* de 44, correspond à une note *étalonnée* de 3 (classe 3). Que signifie cette classe 3 ? Dans cet étalonnage, savoir que le sujet se situe dans la classe 3 nous permet de situer précisément la place du sujet parmi une population de référence (ici : les cadres). Pour effectuer ce positionnement il faut utiliser la seconde colonne du tableau (% théorique). La classe 3 comporte 12,1 % des sujets, on peut donc indiquer ici que 12,1 % des sujets (cadres) obtient une note équivalente, que 10,6 % des sujets (4 % + 6,6 %) obtient une note inférieure et donc que 77,3 % des sujets [100 % – (12,1 % + 10,6 %)][1] obtient une note supérieure. Autrement dit, en première analyse, le sujet qui obtient 43 points ne se situe pas parmi les meilleurs si on le compare aux résultats des sujets exerçant le même métier, c'est même plutôt l'inverse : le score brut de 44 points le situant en classe 3 donc dans les scores plutôt bas.

1. On aurait pu trouver ce même pourcentage, 77,3, en additionnant les autres pourcentages du tableau : 17,5 + 19,6 + 17,5 + 12,1 + 6,6 + 4 = 77,3 %.

Mais si notre sujet qui a obtenu toujours ce score brut de 44 points n'est pas cadre mais est un jeune sans qualification, il convient donc d'utiliser un autre étalonnage : un étalonnage qui correspond à cette population de référence (voir tableau 2.2).

Tableau 2.2
Étalonnage (fictif) du même test de 60 items, en neuf classes normalisées, sujets sans qualifications.

Classes	% théorique	Notes Brutes
1	4	0-9
2	6,6	10-17
3	12,1	18-24
4	17,5	25-32
5	19,6	33-38
6	17,5	39-43
7	12,1	44-48
8	6,6	49
9	4	50 et +

Qu'est-ce qui a changé entre ces deux étalonnages ? Uniquement la troisième colonne, celle qui correspond à la répartition des scores bruts dans la population de référence.

Que peut-on dire de ce score brut de 44 points ? Cette fois, ce score brut de 44 correspond à un très bon score, une note étalonnée de 7, score qui n'est atteint que par environ 23 % des sujets. Le score brut (la performance) n'a pas changé, par contre c'est la population de référence qui est différente et qui explique cette variation du score étalonné.

Cet exemple illustre bien l'objectif de l'étalonnage qui est de transformer le score brut (ici de 44 points) en un score étalonné. Et ce score étalonné dépend, comme nous venons de le voir, de la population de référence. Le niveau de performance observé n'est donc qu'une mesure relative : c'est un indicateur de positionnement du sujet dans une certaine population. Comme nous venons de l'illustrer, une même performance (ici un score brut de 44) sera alors interprétée différemment selon l'étalonnage considéré. Il convient donc toujours de s'interroger sur la population de référence qu'il faut utiliser en fonction de la question posée : s'agit-il de comparer le sujet

aux sujets du même niveau de qualification ? ou aux sujets du même âge ? ou de comparer les performances du sujet à une population générale ?

Mais attention, il existe différents types d'étalonnages : des étalonnages normalisés et des étalonnages par quantilages. Et selon le type d'étalonnage, comme nous allons le voir, l'interprétation de la note étalonnée peut varier.

Plusieurs types d'étalonnage

On distingue deux grandes catégories d'étalonnages : les étalonnages normalisés et les étalonnages par quantilages.

➤ Les étalonnages normalisés

Dans ce type d'étalonnage, chaque classe ne comporte pas le même pourcentage d'individus mais la répartition est effectuée selon la loi Normale. Les limites des classes sont définies ici de manière à respecter cette répartition théorique (courbe de Gauss) : une majorité de sujets dans la classe centrale (qui correspond aux scores proches de la moyenne) et progressivement de manière symétrique de moins en moins de sujets de part et d'autre de cette classe centrale. C'est le type d'étalonnage qui correspond aux étalonnages des tableaux 2.1 et 2.2 de notre exemple de départ : un étalonnage normalisé en 9 classes, avec une majorité de sujets dans la classe centrale (près de 20% des sujets dans cette classe 5 contre 4% dans chaque classe extrême).

Les étalonnages normalisés comportent toujours un nombre impair de classes (5, 7, 9 ou 11 classes) car ils sont centralisés sur une classe centrale, ce qui les différencie des étalonnages par quantilage.

Le tableau 2.3 indique les proportions théoriques de sujets dans les étalonnages normalisés les plus utilisés (on parle de proportion théorique car les effectifs réellement observés peuvent légèrement varier).

Les limites de chaque classe sont déterminées en référence à l'écart type de la distribution. Par exemple, pour un étalonnage normalisé en 5 classes, les bornes correspondent à : –1,5 écart type ; –0,5 ; +0,5 ; +1,5.

À partir des données de ce tableau, il est possible de calculer plus précisément la position d'un sujet particulier dans une population de référence.

Pour interpréter plus facilement ces positionnements, certains étalonnages procèdent à des regroupements de notes étalonnées avec des catégories en nombre plus limités sous la forme de codage : – –, –, 0, +, ++ .

Tableau 2.3

Répartition théorique des étalonnages normalisés.

5 classes	1 6,7 %		2 24,2 %		3 Classe centrale 38,2 %		4 24,2 %		5 6,7 %		
7 classes	1 4,8 %		2 11,1 %		3 21,2 %	4 Classe centrale 25,8 %		5 21,2 %		6 11,1 %	7 4,8 %
9 classes	1 4 %	2 6,6 %	3 12,1 %	4 17,5 %	5 Classe centrale 19,6 %		6 17,5 %	7 12,1 %	8 6,6 %	9 4 %	
11 classes	1 3,6 %	2 4,5 %	3 7,7 %	4 11,6 %	5 14,6 %	6 Classe centrale 16 %	7 14,6 %	8 11,6 %	9 7,7 %	10 4,5 %	11 3,6 %

Tableau 2.4

Codage des scores à partir d'un étalonnage en 9 classes normalisé.

Notes étalonnées								
– –	–		0			+		++
1	2	3	4	5	6	7	8	9
4,4 %	6,6 %	12,1 %	17,5 %	19,6 %	17,5 %	12,1 %	6,6 %	4,4 %

En plus du positionnement classique en 9 classes, on dispose ici d'un autre codage de la performance du sujet. Il s'agit presque d'un autre étalonnage, qui repère ici 5 niveaux de réussite (de la classe - - à la classe ++) :

– une réussite *moyenne*, la classe centrale, la classe 0, regroupant plus de la moitié des sujets (54,6 %)[1] ;
– une réussite *au-dessus de la moyenne*, la classe +, regroupant elle un peu moins de 20 % des sujets (18,7 %)[2] ;
– symétriquement, une réussite *en dessous de la moyenne*, la classe – ;
– et enfin, les deux classes extrêmes, qui correspondent à des niveaux de réussite très *inférieurs* (classes – –), ou très supérieurs (classe ++), à la moyenne.

1. 17,5 + 19,6 + 17,5 = 54,6 %
2. 12,1 + 6,6 = 18,7 %

Attention ici à ne pas faire de faux-sens un score 0 correspond ici à un score moyen, ou autour de la moyenne, et non pas à un score nul !

➤ Les étalonnages par quantilages

Dans ce type d'étalonnage chaque catégorie étalonnée (chaque classe) comporte le même pourcentage d'individus. On utilise habituellement des étalonnages en 10 classes : les décilages. Dans un étalonnage de type décilage, chaque classe comporte alors 10 % des individus. Cet étalonnage est très simple à réaliser. Il est également simple à utiliser car on peut situer très rapidement la position du sujet. Prenons un exemple d'un test comportant 40 items (voir tableau 2.5).

Tableau 2.5

Exemple d'étalonnage en déciles pour un test de 40 items.

Notes étalonnées	1	2	3	4	5	6	7	8	9	10
Notes brutes	0–14	15–19	20–22	23–24	25	26–27	28–29	30–32	33–34	35 et +

Un sujet qui obtient un score brut de 28 se situe alors dans la classe 7 de cet étalonnage : on sait que 10 % des sujets ont un score équivalent au sien, que seulement 30 % des sujets ont un score supérieur au sien et 60 % ont un score inférieur.

On trouve également des références en terme de percentile ou de rang percentile. Dans un étalonnage en rang centile il y a 100 classes, chacune comportant 1 % des sujets (classe 1 à classe 100). Ce type d'étalonnage va situer le sujet par son rang : le 50^{e} centile correspondant, par exemple, à une position médiane : 50 % des sujets ont un score inférieur ou égal et 50 % des sujets ont un score supérieur. C'est pour cette raison que le 50^{e} centile correspond à la médiane de la distribution.

De la même manière, dire que le sujet se situe au percentile 75 c'est dire qu'il occupe la 75^{e} place sur 100, à partir du score le plus bas. Autrement dit, 75 sujets (soit 75 %) ont un score inférieur ou égal au sien et 25 (soit 25 %) un score supérieur au sien.

Ce type d'étalonnage permet de situer globalement le sujet parmi les quartiles[1] de la distribution d'étalonnage : le premier quartile correspondant

1. Les quartiles permettent de diviser une distribution en quatre classes d'effectifs égaux : on trouve 25 % des sujets entre chaque quartile.

au percentile 25, le second quartile à la médiane, le troisième quartile au centile 75.

Certains tests, comme les matrices de Raven, proposent de tels étalonnages en rang centiles (ou percentile) : voir un exemple sur le tableau 2.6.

Tableau 2.6
Extrait de l'étalonnage INETOP en rang percentile pour le test SPM et pour des élèves de 3e de collège.

	Percentile						
	5	10	25	50	75	90	95
Score brut	36	38	42	46	48	51	53

Si un élève de troisième obtient un score brut de 42 points, il se situe alors au 25e centile, c'est-à-dire en 25e position par rapport au score le plus bas : 25 % des sujets ont donc un score inférieur ou égal au sien et 75 % obtiennent un score supérieur au sien. Il est donc situé juste à la limite du premier quartile.

Si un élève obtient un score brut de 44, il se situe donc entre le 25e centile et le 50e centile, c'est-à-dire entre le premier et le deuxième quartile. Son score le situe donc en dessous de la moyenne.

Pour qualifier la performance du sujet il est possible ici aussi à procéder à une catégorisation des scores. Un exemple d'une telle catégorisation figure dans le manuel du SPM (tableau 2.7).

On peut remarquer sur ce tableau que les scores peuvent être catégoriés de la classe I à la classe V, la classe centrale III regroupant 50 % des sujets, et que les classes extrêmes (classe I et classe V) regroupant chacune 5 % des sujets.

➤ Avantages et inconvénients de ces deux types d'étalonnage

L'avantage principal des étalonnages normalisés c'est qu'ils différencient de manière plus fine les scores extrêmes. Mais ils sont moins discriminants sur les scores moyens. Par contre, c'est l'inverse pour les étalonnages de type quantilages. Mais, généralement, le praticien ne peut pas choisir entre ces deux types d'étalonnages car les étalonnages fournis avec un test diffèrent sur la composition des échantillons d'étalonnage (on dispose alors de plusieurs populations de comparaison) mais pas sur le type d'étalonnage.

Tableau 2.7
Catégorisation des scores au test SPM (d'après le manuel SPM, section 3, p. 51).

Catégorisation	Conditions
Classe I : « capacité intellectuelle supérieure »	si le score atteint ou dépasse le centile 95 des sujets de son groupe d'âge
Classe II : « capacité intellectuelle nettement au-dessus de la moyenne »	si le score atteint ou dépasse le centile 75 (classe II+ si le score atteint ou dépasse le centile 90)
Classe III : « capacité intellectuelle moyenne »	si le score se situe entre les centiles 25 et 75 (plus de 50 : classe III+ ; moins de 50 : classe III –)
Classe IV : « capacité intellectuelle nettement inférieure à la moyenne »	si le score se situe au centile 25 ou au dessous (classe IV– si le score se situe au centile 10 ou en dessous)
Classe V : « déficience intellectuelle »	si le score se situe au centile 5 ou au dessous

➤ Les étalonnages de type Q.I.

Les étalonnages de type Q.I. sont en fait des étalonnages normalisés avec une moyenne de 100 et un écart type de 15.

Nous avons vu précédemment (dans le chapitre 1) que la notion de Q.I. est héritée de la notion d'âge mental proposée par Binet.

À l'origine le Q.I. – Quotient Intellectuel – est bien un quotient, c'est-à-dire le résultat d'une division. C'est Stern en 1913 qui propose ce calcul de Q.I. comme étant le rapport entre l'âge mental d'un enfant (évalué par un test) et son âge réel (son âge biologique) :

$$\text{Q.I.} = \frac{\text{âge mental} \times 100}{\text{âge réel}}$$

Avec cette formule, si l'âge mental correspond à l'âge réel, l'enfant a donc un Q.I. de 100. Si son âge mental est supérieur à son âge réel son Q.I. sera supérieur à 100. Et inversement, si son âge mental est inférieur à son âge réel son Q.I. sera inférieur à 100.

Mais cet indicateur présentait des limites, comme par exemple celle de ne pas être applicable à des adultes. Dans les épreuves créées ensuite, comme par exemple les échelles de Wechsler, la notion de Q.I. va être reprise mais elle ne fera plus référence à cette notion d'âge mental, la notion de Q.I. indiquera uniquement un indicateur de positionnement dans une population de référence (principe de l'étalonnage).

Avec le Q.I., la population de référence est toujours la population du même âge, que ce soit pour les enfants (par exemple le WISC-III ou le WISC-IV) ou pour les adultes (la WAIS-III).

Mais Wechsler, par construction, conserve la référence à 100, qui constitue alors le score moyen. L'autre indicateur important de cette distribution de Q.I. est l'écart type : il est ici de 15. Comme la distribution suit une loi normale, ces deux valeurs (moyenne et écart type) nous permettent de calculer des répartitions de sujets.

La figure 2.2 nous permet de visualiser ces répartitions théoriques pour différents types de scores.

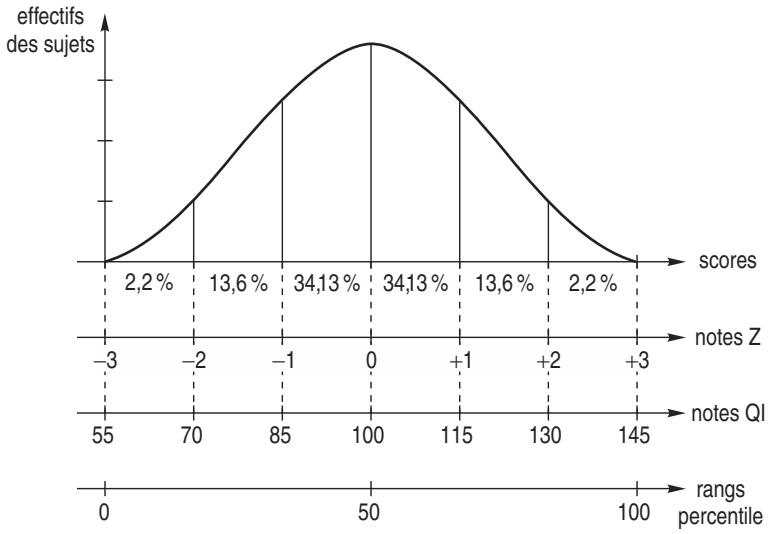

Figure 2.2

Répartition théoriques des sujets en fonction des différents types de scores.

La figure 2.2 indique les proportions de sujets relatives à différents indicateurs de test. La première ligne concerne les notes z, notes centrées réduites (moyenne de 0 et écart type de 1). La deuxième ligne concerne les notes de Q.I. de moyenne 100 et d'écart type 15. Enfin, figurent en quatrième ligne les rangs percentiles.

Les proportions indiquées permettent de mieux situer les performances des sujets. Par exemple, sur une échelle de QI, entre 85 et 100, soit un écart type en dessous de la moyenne, se situent 34,13 % des sujets. Comme la courbe normale est symétrique, on peut donc calculer très rapidement la proportion de sujets situés dans l'intervalle [moyenne – un écart type ;

moyenne + un écart type], donc ici entre 85[1] et 115[2] : environ 68,3 % des sujets ont ainsi un Q.I. compris entre 85 et 115.

De même, toujours pour les scores de Q.I., la figure F1 nous indique également que :

– 15,73 % des sujets a un Q.I. supérieur à 115 (moyenne plus un écart type), et la même proportion a un Q.I. inférieur à 85 (moyenne moins un écart type) ;
– 2,14 % des sujets se situent au-delà de 130 (moyenne plus deux écarts type), et la même proportion a un Q.I. inférieur à 70 (moyenne plus deux écarts type).

À partir de cette répartition des Q.I. dans la population, il est possible, là encore, d'établir des catégorisations. Mais cette catégorisation peut varier, d'une part selon l'époque, d'autre part selon l'auteur de cette catégorisation.

Par exemple, dans la catégorisation proposée en 1928 par Levine et Marks (cité par Bernaud, 2000a) les catégories allaient de « idiot » (pour un Q.I. situé entre 0 et 24) à « précoce » (score supérieur à 175) et dans la catégorisation de Terman (cité par Wechsler, 1956, p. 47) de « débilité mentale caractérisée » (pour un Q.I. au-dessous de 70) à « génie ou sujet proche du génie » (Q.I. supérieur à 140). On peut également observer des différences dans la dénomination des catégories entre deux versions d'une même épreuve, comme par exemple entre les versions WAIS-R et WAIS-III : on peut remarquer que, si les bornes n'ont pas changé, la dénomination des classes elle a été modifiée (voir tableau 2.8).

Plusieurs remarques :

- Un Q.I égal ou supérieur à 130, qui correspond au score d'environ 2 % des sujets, est qualifié de « très supérieur » et correspond généralement à l'un des critères pour diagnostiquer une « précocité intellectuelle » (sur cette problématique de précocité intellectuelle voir le numéro spécial de la revue *Psychologie Française* de 2004 coordonné par Lautrey) ;
- L'une des modifications entre les deux catégorisations concerne les deux catégories de part et d'autre de la moyenne : *Normal Fort* devient *Moyen Supérieur* et, symétriquement, *Normal Faible* devient *Moyen Inférieur* ;
- La seconde modification, sans doute la plus visible, concerne la catégorie la plus basse : *Déficient mental* devenant *Très faible*. L'explication

1. 100 –15 = 85
2. 100 + 15 = 115

Tableau 2.8
Classification des Q.I. aux échelles de Wechsler pour adultes (d'après Wechsler, 1989, p. 24 et Wechsler, 2000, p. 280).

Q.I	Pourcentage théorique de sujets	Classification WAIS-R (1989)	Classification WAIS-III (1997)
130 et plus	2,2 %	Très supérieur	Très supérieur
120–129	6,7 %	Supérieur	Supérieur
110–119	16,1 %	Normal fort	Moyen supérieur
90–109	50 %	Moyen	Moyen
80–89	16,1 %	Normal faible	Moyen inférieur
70–79	6,7 %	Limite	Limite
69 et moins	2,2 %	Déficient mental	Très faible

est donnée dans le manuel de la WAIS-III : les auteurs justifient ce changement afin qu'un Q.I. très faible ne soit pas considéré comme reflétant obligatoirement une déficience mentale (manuel WAIS-III, p. 280).

D'autres classifications existent comme par exemple celle de l'O.M.S (Organisation Mondiale de la Santé) qui distingue *retard mental léger* (Q.I. compris entre 50 et 70) et *retard mental moyen* (Q.I. compris entre 35 et 49) (voir sur ce point Lathoud, 1997).

Enfin, et pour terminer avec les étalonnages de type Q.I., il faut rappeler que le Q.I. est un indicateur du positionnement du sujet dans sa classe d'âge. Il ne s'agit donc pas d'une mesure absolue des capacités cognitives du sujet dans le sens où, par exemple, un enfant de 12 ans qui présente un Q.I. de 115 a, en réalité, des performances inférieures à celles d'un enfant plus âgé qui présente pour tant ce même score de 115. Par contre, ce que veut dire ce score de 115 c'est que ces deux enfants se situent de la même manière *dans leur population respective*, et plus précisément, à un écart type au-dessus de la moyenne. Cette remarque vaut également pour les adultes : un indicateur de type Q.I. est un positionnement dans une classe d'âge, même pour des adultes.

Signalons qu'il existe encore d'autres étalonnages, mais beaucoup moins utilisés en France, comme par exemple des étalonnages utilisant les scores T (avec une distribution de moyenne égale à 50, et un écart type de 10) et les stanines (moyenne de 5 et écart type de 2).

Conclusion sur la notion d'étalonnage

Nous voudrions ici conclure sur trois aspects importants : la fiabilité des étalonnages, le type d'étalonnage choisi et la prise en compte de l'erreur de mesure.

Il convient, avant toute passation de s'assurer de la qualité des étalonnages accompagnant le test que l'on veut utiliser. Le psychologue sera attentif aux caractéristiques des populations d'étalonnage afin, d'une part, de juger de la fiabilité des normes, et, d'autre part, de vérifier qu'au moins un des étalonnages proposés correspond aux caractéristiques du sujet devant passer le test (condition minimum de la comparabilité des résultats).

Concernant le premier point, l'analyse de la fiabilité des normes, il faut particulièrement étudier :

1. *Le nombre de sujets composant l'étalonnage*.
   Le nombre minimum de sujets dépend du type d'échantillonnage choisi (Laveault et Grégoire, 1997) mais on peut retenir qu'un étalonnage comportant moins d'une centaine de sujets ne présente pas une fiabilité satisfaisante.

2. *La sélection de ces sujets.*
   L'échantillon d'étalonnage doit être représentatif de la population qu'il est censé représenter. Selon les cas, cette population est plus ou moins vaste. Par exemple, dans le cas des échelles de Q.I., la population de référence est constituée par les sujets de même âge. L'échantillon d'étalonnage doit donc comporter, pour chaque niveau d'âge, les mêmes caractéristiques que la population parente (répartition des sexes, des CSP, des niveaux scolaires...). S'il s'agit d'un étalonnage spécifique, comme par exemple un étalonnage par niveau d'étude ou par profession, la population parente est alors plus réduite mais l'échantillon d'étalonnage doit toujours en être un bon représentant.

3. *la date de l'étalonnage.*
   Un étalonnage trop ancien ne présentera pas une référence fiable. On peut retenir comme seuil une dizaine d'années : un test qui présenterait des étalonnages datant de plus de 15 ans sera à utiliser avec prudence. En effet, une augmentation progressive des scores aux tests d'intelligence a été observée au cours du XX^{e} siècle. Cet effet, connu sous le terme d'*effet Flynn* a été présenté dans le chapitre 1). Utiliser un étalonnage trop ancien peut avoir alors comme conséquence de surestimer les capacités du sujet par rapport à sa population de référence.

Il faut indiquer ici que les étalonnages figurent dans le manuel du test mais qu'il existe souvent des étalonnages complémentaires, sous formes d'annexes, qui ont été réalisés après le manuel et que les éditeurs de tests peuvent fournir. Nous encourageons donc le praticien à s'informer auprès des éditeurs des étalonnages les plus récents disponibles pour le test qu'il souhaite utiliser.

Il est aussi très important de repérer le type d'étalonnage que l'on va utiliser de manière à situer sans erreur la position du sujet dans la population de référence. Par exemple, un score étalonné de 3 dans un étalonnage de type décilage, ne correspond pas à un score étalonné de 3 dans un étalonnage normalisé en 5 classes. Dans le premier cas, seulement 20 % des sujets obtiennent un score inférieur, la performance se situe donc largement en dessous de la moyenne, tandis que, dans le second cas, ce score correspond à un score moyen (voir tableau 2.3).

Il est regrettable d'ailleurs que, parfois, le type d'étalonnage ne soit pas clairement indiqué dans le manuel. En l'absence d'information explicite sur ce point, il faut se rappeler que les étalonnages par quantilages sont le plus souvent des décilages (10 classes), et que les étalonnages normalisés comportent toujours un nombre impair de classes (5, 7, 9 ou 11 classes). En cas de doute, le praticien doit demander une clarification auprès de l'éditeur du test afin d'éviter toute erreur d'interprétation à ce niveau.

Enfin, il faut également prendre en compte la notion d'erreur de mesure et la distinction entre score vrai et score observé**.** L'idéal est de calculer un intervalle de confiance pour situer plus précisément le score vrai du sujet à partir du score observé, mais, à défaut, il faut au moins se rappeler que si le score brut du sujet est situé à proximité d'une des bornes de la classe étalonnée, le score vrai du sujet pourrait se situer de l'autre côté de cette borne. Il convient alors de nuancer l'interprétation du score observé.

Cette capacité d'analyse critique des étalonnages fournis, et plus globalement d'analyse de toute information figurant dans le manuel, cette capacité d'interprétation d'un score observé, de recul par rapport à une mesure, représente l'un des fondements d'une pratique professionnelle de psychologue. Nous allons développer ce point dans la partie suivante.

8. Comment évaluer un test ?

Un test doit toujours être accompagné d'une documentation technique, prenant le plus souvent la forme d'un manuel dans lequel figurent les informations nécessaires à la passation du test (consignes, modalités de cotation, étalonnages...). Mais ce manuel doit également comporter toutes les informations relatives aux différentes phases d'élaboration du test et de sa validation (création des items, sélection des items, analyse des biais, études de fidélité, de validité...). C'est en prenant connaissance de toutes ces informations qu'un psychologue pourra juger, d'une part, de la pertinence d'utiliser ce test par rapport à la situation dans laquelle il se trouve, d'autre part, de la fiabilité du test. En effet le psychologue ne doit utiliser que des outils dont la fiabilité a été évaluée :

> « Les techniques utilisées par le psychologue pour l'évaluation, à des fins directes de diagnostic, d'orientation et de sélection, doivent avoir été scientifiquement validées. » (Code de déontologie des psychologues article 18).

Mais le psychologue ne doit pas s'en tenir à un niveau trop superficiel d'information sur le test, il doit faire preuve de professionnalisme en recherchant dans la documentation technique qui accompagne le test (le ou les manuels, les annexes...) les éléments qui doivent témoigner de la fiabilité de ce test.

Ce sont ces éléments de validation, d'évaluation de la qualité de la mesure qui distingue l'évaluation psychologique d'autres pratiques à visée évaluative (comme, par exemple, la graphologie...).

C'est d'ailleurs cette démarche de recherche des éléments de validation qui est préconisée dans les *recommandations internationales sur l'utilisation des tests*, en particulier dans la section 2 :

« Assurer une pratique correcte dans l'utilisation des tests » (pages 19 à 24). Quelques citations : « Se garder de l'utilisation de tests qui ont une documentation technique inadaptée ou peu claire » ; « se garder de porter un jugement sur un test seulement sur la base de sa validité apparente, des témoignages des utilisateurs, ou du conseil de personnes qui y ont des intérêts commerciaux » ; apprécier « la précision de la mesure », « la fidélité », « la validité », « l'absence de biais »... ; s'assurer que « les tests ne sont pas biaisés et sont adaptés pour les différents groupes qui vont être testés. »...

Nous retrouvons, sous une autre forme, certaines de ces recommandations dans les questions formulées par Rolland (2001) concernant l'analyse de la fiabilité d'un test :

1. Quelle est la précision de ce test ? Quelle est sa marge d'erreur ?
2. Le test mesure-t-il réellement ce qu'il est censé mesurer ?
3. Les informations recueillies par ce test sont-elles pertinentes (utiles) pour l'objectif suivi ?
4. Les scores fournis par ce test permettent-ils de bien différencier les sujets ?

Nous reprendrons ces questions qui nous permettent de synthétiser les principaux points que nous venons d'aborder dans ce chapitre.

1. *Quelle est la précision de ce test ? Quelle est sa marge d'erreur ?*
   Ces questions renvoient à la notion de fidélité. Il convient d'analyser deux critères : la stabilité et l'homogénéité de la mesure. Un test fiable doit présenter des indices de stabilité et d'homogénéité d'au minimum de .70.
2. *Le test mesure-t-il réellement ce qu'il est censé mesurer ?*
   Cette question fait référence à la notion de validité que l'on peut en particulier apprécier sous l'angle de la validité de structure et de la validité convergente.
   - La validité *de structure* : la corrélation d'un item sur une dimension (on parle de saturation de l'item sur la dimension) doit être de .30 au minimum et la liaison avec la dimension attendue doit être plus élevée que la liaison avec une autre dimension. Les dimensions ne doivent pas être trop liées les unes aux autres car des dimensions fortement corrélées sont redondantes.
   - La validité *convergente* (analyse des liaisons entre 2 épreuves censées évaluer la même dimension) : une corrélation autour de .70 est considérée comme satisfaisante, avec une valeur minimale de .40.
3. *Les informations recueillies par ce test sont-elles pertinentes (utiles) pour l'objectif suivi ?*
   Cette question se réfère à la validité prédictive, ou validité critérielle, de l'épreuve. La question ici est de savoir ce que permet de prédire le test : réussite scolaire pour les enfants et les adolescents, réussite en formation pour les adultes, réussite professionnelle, par exemple. À titre de référence utile pour l'interprétation des coefficients de validité, nous retiendrons que les tests d'aptitude cognitive corrèlent en moyenne à .50

avec des critères de réussite scolaire et de performances professionnelles. Cet aspect sera abordé plus en détail dans le chapitre 8.

4. *Les scores fournis par ce test permettent-ils de bien différencier les sujets ?*
   Cette dernière question, fait référence à la sensibilité du test et à son adéquation à la personne évaluée. Rolland (2001) rappelle que la distribution des scores doit suivre une courbe de Gauss et qu'il faut analyser ici la qualité des étalonnages fournis.
   Si l'étalonnage est un étalonnage *général* il doit être représentatif des caractéristiques de la population. Il faut donc s'assurer des modalités de sélection des échantillons d'étalonnage : sexe, âge, niveau d'étude, profession... (démarche identique à celle utilisée dans les sondages pour créer un échantillon représentatif d'une population). Dans ce cas l'échantillon d'étalonnage doit comporter au minimum 500 sujets.
   Si l'étalonnage est *spécifique*, il correspond alors à une catégorie de la population (étalonnage par âge, par profession, par niveau d'étude...) il doit comporter au minimum 200 sujets.
   Enfin, Rolland nous indique que des tests dont les normes (les étalonnages) datent de plus de 10 ans ne devraient pas être utilisés en raison de l'effet Flynn.

Rolland précise que les valeurs seuils indiquées pour les différentes corrélations doivent être prises avec souplesse car il convient de prendre également en compte, dans l'analyse de la fiabilité d'un test, l'ensemble des éléments fournis par le manuel. Ainsi, vaut-il souvent mieux utiliser un test qui présente des valeurs un peu plus faibles qu'attendues sur ces indices, qu'utiliser un test pour lequel nous n'aurions pas de données statistiques concernant sa validation...

Les indicateurs qui viennent d'être évoqués comme pertinents pour évaluer un test font directement référence aux notions de base de la psychométrie présentées dans ce chapitre. Il est donc indispensable que le praticien les maîtrise afin de pouvoir analyser de façon critique les outils qu'il utilise. Cette analyse est à mener à partir des informations présentes dans les manuels des tests.

Une lecture attentive des manuels doit également permettre d'améliorer l'interprétation des indicateurs de la performance d'un sujet recueillis par le test. Nous en donnerons deux exemples, l'un relatif aux sous scores du test NNAT et l'autre aux conditions de passation des matrices de Raven (*cf.* encadré). Ces deux exemples illustrent bien tout l'intérêt d'une lecture approfondie des différents documents accompagnant le test car la qualité de

l'interprétation des scores du sujet va dépendre en grande partie de la prise en compte de ces informations.

On peut remarquer ici que la nature et la qualité de ces informations diffèrent selon les tests : certains proposent des manuels riches d'information, avec parfois plusieurs manuels pour une même épreuve (en distinguant par exemple, un manuel d'utilisation et un manuel d'interprétation), tandis que d'autres tests ne proposent qu'un manuel sommaire. Le choix du test par le praticien doit prendre en compte cet élément.

Exemples

– **Dans le test NNAT** (qui sera présenté dans le chapitre 4) il est possible de calculer, en plus du score total, quatre sous-scores qui correspondent à quatre types de raisonnement. Mais attention ici à l'analyse de ces sous-scores car ils ne présentent pas la même fiabilité que le score total :

> « Les sous-scores devront être interprétés que de façon qualitative à partir des notes brutes. » (Manuel du NNAT, p. 12)

Deux raisons sont avancées dans le manuel : le nombre d'items de chaque sous-score et la validation de ces sous-scores. En effet, d'une part, le nombre d'items est variable selon les sous-scores et reste un peu trop faible pour garantir un bon niveau de fiabilité de la mesure, d'autre part, les analyses statistiques rendent discutable la distinction même de ces quatre types de raisonnement (manuel NNAT, p. 49). Pour ces deux raisons les auteurs indiquent qu'il n'a pas été possible de réaliser un étalonnage spécifique de chaque sous score ce qui, selon nous, retire alors beaucoup d'intérêt au calcul de ces sous-scores.

On peut voir clairement ici qu'une démarche automatique de calcul, et d'interprétation, de ces sous-scores aboutirait alors à des interprétations qui ne reposeraient pas sur des éléments suffisamment fiables. Par contre, une lecture attentive des informations du manuel du NNAT devrait aboutir à relativiser ces indicateurs de sous-scores et à les interpréter avec prudence.

– **Pour le test SPM** (PM38) de Raven (également présenté dans le chapitre 4), il est indiqué dans le manuel qu'il est plus fiable de faire passer l'épreuve en temps libre en raison de l'effet possible du style de réponse du sujet. En effet, certains sujets peuvent « sauter » les items difficiles et répondre d'abord aux items les plus faciles, quitte à effectuer ensuite un retour en arrière s'ils ont du temps, tandis que d'autres sujets vont

prendre du temps afin de rechercher la réponse à ces items difficiles, mais du coup, n'auront peut-être pas le temps d'aborder des items plus faciles, situés vers la dernière partie de l'épreuve (Manuel Matrice de Raven, Section 1, p. 66). Nous reviendrons plus loin, lors de la présentation de ce test, sur ces styles de réponse. Si le praticien utilise ce test en temps limité il est donc souhaitable qu'il analyse le patron de réponse du sujet afin de s'assurer de l'absence d'une stratégie de ce type. Le praticien connaît-il toujours l'existence de ce biais possible ? S'il n'a pas pris connaissance de ces éléments qui figurent dans le manuel général des Matrices de Raven, il est fort probable qu'il ignore cette possibilité de biais.

9. Les évolutions des modèles psychométriques

Comme nous l'avons indiqué au tout début de ce chapitre, la quasi-totalité des tests utilisés actuellement en France reposent sur la théorie classique des tests, fondée sur la notion de score vrai et d'erreur de mesure. D'autres modèles de mesure existent que nous ne pouvons pas présenter ici car ils dépassent largement l'objectif de cet ouvrage. Les lecteurs intéressés par une présentation de ces différents modèles de mesure pourront consulter les ouvrages spécialisés comme celui de Dickes *et al.* (1994), ou celui de Laveault et Grégoire (2002).

Néanmoins, il nous a semblé intéressant d'aborder ici l'un de ces modèles : le modèle de Réponse à l'Item (M.R.I). En effet, l'utilisation de ce modèle, ou plutôt de ces modèles MRI (nous verrons qu'il existe plusieurs modèles MRI), ou modèles I.R.T[1], est croissante, au moins au niveau international et dans le domaine de l'évaluation des connaissances scolaires, et il est fort probable que d'ici quelques années des tests reposant sur ces modèles MRI soient disponibles en France. Il est donc important de connaître les principes de base de ces modèles de mesure. Avec les modèles M.R.I il s'agit d'un autre modèle de la mesure, un modèle probabiliste dans lequel certaines notions classiques de psychométrie, comme par exemple les notions de difficulté de l'item, d'étalonnage, ou encore de score du sujet, vont être profondément modifiées.

1. En anglais on utilise le terme IRT pour *Item Response Theorie*. Mais le terme de modèle semble plus approprié (Vrignaud, 1996).

Notre objectif ici est de donner une information minimale sur ces modèles MRI, accessible à tout psychologue. Pour cette raison nous éviterons l'utilisation d'équations et de formules mathématiques, qui servent à l'estimation des paramètres des modèles (voir plus loin) que le lecteur pourra trouver dans les ouvrages spécialisés de psychométrie (déjà cités) ou dans des publications traitant spécifiquement de ces modèles (voir par exemple le numéro spécial de la revue *Psychologie et Psychométrie* coordonné par Juhel en 1999 ; l'article de Vrignaud de 1996 ; ou encore l'annexe consacrée aux MRI dans l'ouvrage de Reuchlin de 1997).

Présentation générale de l'approche des modèles MRI

Dans l'approche classique de la mesure les principaux indices psychométriques qui vont caractériser l'épreuve (indices de difficulté des items, étalonnages...) vont dépendre de l'échantillon de sujets utilisé. Par exemple, dans un test de facteur g comme le SPM de Raven qui peut être utilisé sur des populations de niveaux très différents (collégiens, lycéens, adultes...), un même item sera considéré comme difficile pour des collégiens, mais comme facile (ou plus facile) pour des étudiants. Autrement dit, on ne connaît pas le niveau *absolu* de difficulté d'un item car il s'agit toujours un niveau *relatif* de difficulté qui va dépendre directement du niveau des sujets ayant passé le test.

Parallèlement, pour un sujet confronté aux items d'un test, le niveau de sa performance (le score observé) ne peut s'interpréter que par comparaison avec le niveau de réussite d'autres sujets présentant les mêmes caractéristiques (par exemple sujets de même âge) : c'est le principe de l'étalonnage. Il s'agit là aussi de mesure relative.

En d'autres termes, les caractéristiques des items (en particulier leur niveau de difficulté) sont dépendantes des caractéristiques des sujets mais les caractéristiques des sujets (en particulier leur niveau de réussite, c'est-à-dire leurs scores) sont dépendantes des caractéristiques des items.

Dans la théorie classique des tests, il y a donc interdépendance entre caractéristiques des items et caractéristiques des sujets.

C'est l'une des différences principales avec les modèles MRI dont l'objectif principal est de permettre une évaluation indépendante de ces deux séries d'éléments : caractéristiques des sujets et caractéristiques des items. Avec ces modèles il devient donc possible, après une phase de calibrage des items (voir

plus loin), d'évaluer le niveau de performance du sujet, quels que soient les niveaux de difficulté des items (donc le test) qu'il aura passé.

Les modèles MRI sont des modèles probabilistes : des modèles dans lesquels on cherche à estimer des probabilités de réussite. La principale propriété des MRI est de placer les difficultés des items et les compétences des sujets sur une même dimension : le trait latent θ (*theta*). Ce trait latent représente la variable évaluée qui peut être, selon les cas, une aptitude cognitive, une compétence scolaire, voire un trait de personnalité...

L'un des postulats de base est le suivant : les différences de réussite entre les sujets s'expliquent par ce trait latent θ et uniquement par celui-ci.

Dans les tests d'intelligence, le trait latent représente donc l'intelligence telle qu'elle est évaluée par le test. Dans ce cadre on utilise préférentiellement le terme compétence pour désigner ce trait latent. . Chaque sujet peut donc être caractérisé sur ce trait, par son niveau de compétence, et parallèlement, chaque item peut être situé sur cette même échelle θ par son niveau de difficulté.

Ainsi, plus le sujet se trouve à un niveau élevé sur ce trait, plus son score θ est élevé, et plus sa probabilité de réussir un item particulier augmente (modèle probabiliste du niveau de compétence du sujet). Parallèlement, plus l'item se situe à un niveau élevé sur ce même trait θ, plus son niveau de difficulté augmente et donc, plus sa probabilité d'être réussi diminue (modèle probabiliste du niveau de difficulté de l'item).

Chaque sujet a, selon son niveau de compétence estimé, une certaine probabilité de réussir un item donné dont la difficulté a été estimée sur cette même échelle de compétence θ. Les modèles de réponse à l'item (MRI) visent à prédire la probabilité qu'un individu I fournisse une bonne réponse à un item i.

Pour bien comprendre la logique de ces modèles MRI, il faut prendre en compte qu'il s'agit d'une modélisation des probabilités de réussite, modélisation effectuée à partir des observations sur la fréquence des bonnes réponses (réussite) observées par un groupe de sujets à une série d'items constituant un test.

Exemple

Soit un test **X** passé par un ensemble de sujets. Ce test comporte 60 items, le score total de chaque sujet se situe donc entre 0 et 60. Nous pouvons représenter sur une figure les fréquences de réussite à un item **A** du test en fonction du score total au test **X**.

On obtient généralement la tendance suivante : plus les sujets ont un score total élevé au test **X**, plus la fréquence de réussite à cet item **A** est élevée.

Inversement, plus les sujets ont un score total faible au test, plus la fréquence de réussite à un item donné diminue.

La figure 2.3 permet de visualiser cette relation : le score total est porté en abscisse, la fréquence de réussite à l'item **A** étant en ordonnée.

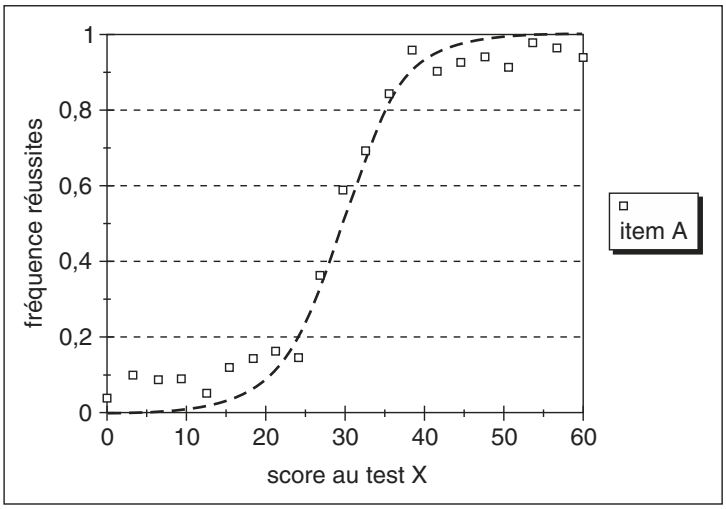

Figure 2.3
Fréquences de réussites observées à l'item A en fonction du score total obtenu à un test X (d'après Vrignaud, 1996, p. 8).

Sur la figure 2.3 chaque carré représente la fréquence de réussite observée pour un score total donné. Par exemple, les sujets ayant un score total inférieur à 25 points (donc les sujets de bas niveau ici) ont une fréquence de réussite à l'item *A* assez faible, inférieure à .20 (soit 20 % de réussite pour ces sujets à cet item *A*). Par contre, les sujets de bon niveau, ayant un score total au test *X* supérieur à 40 points, réussissent beaucoup plus fréquemment cet item *A* avec une fréquence de réussite ici de l'ordre de .90 (soit environ 90 % de réussite pour ces sujets).

Cet exemple illustre bien le fait que la fréquence de réussite observée à un item est faible pour les sujets de bas niveau, puis cette fréquence augmente rapidement pour les sujets de niveau moyen (ici autour du score de 30 points) et atteint un plateau, proche de la fréquence de 1, pour les sujets de niveau élevé.

Le principe fondamental des modèles MRI est de proposer un modèle mathématique permettant de modéliser cette forme de relation. Le modèle

mathématique retenu, le plus représentatif de la relation illustrée par les carrés dans la figure 2.3, est la fonction logistique. La courbe en pointillé de la figure 2.3 représente une telle modélisation. Cette courbe est appelée la courbe caractéristique de l'item (CCI). Une telle courbe est présentée dans la figure 2.4.

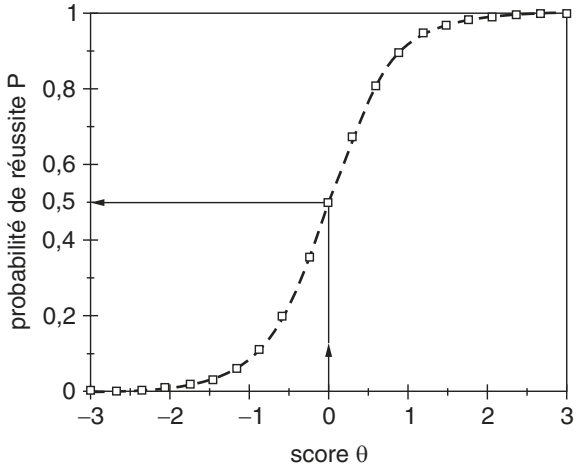

Figure 2.4
Exemple de courbe caractéristique d'un item (CCI).

La figure 2.4 représente bien une modélisation mathématique de la relation représentée dans la figure 2.3. On remarquera que maintenant en ordonné figurent les probabilités de réussite P estimées (et non plus les fréquences de réussite observées) et en abscisse le score θ correspondant au niveau de compétence des sujets (et non plus le score total au test).

La probabilité de réussite P varie donc, comme toute probabilité, de 0 à 1, le niveau de compétence θ des sujets variant lui d'environ[1] –3 à +3, avec une moyenne de 0.

Les courbes CCI de tous les items du test définissent les caractéristiques de ces items. Elles sont estimées par des logiciels spécialisés à partir de données réelles de passation : c'est la phase dite de calibrage des items. Chaque item sera alors caractérisé par différents paramètres (voir plus loin) dont le principal est son niveau de difficulté exprimé sur l'échelle *theta* θ.

1. En réalité, comme il s'agit d'un modèle probabiliste, le score θ peut théoriquement varier de moins l'infini à plus l'infini, mais on estime que 99,8 % des sujets se situent entre -3 et +3 (Laveault et Grégoire, 1997, p. 292).

Lorsqu'un regroupement d'items calibrés est réalisé pour élaborer un test, les sujets peuvent alors être également caractérisés par un score θ représentant leur niveau de compétence.

Par convention, on considère qu'un score θ de 0 correspond au score moyen, un score positif correspond alors à un score au-dessus de la moyenne, un score négatif à un score en dessous de la moyenne.

Comme nous venons de l'indiquer, le niveau de difficulté ne représente qu'une des caractéristiques possibles pour définir un item. Nous allons maintenant présenter succinctement les trois principaux modèles MRI qui diffèrent sur le nombre de caractéristiques (ou paramètres) pris en compte.

Les trois modèles MRI

➤ Le modèle à un paramètre (indicateur b)

Dans ce modèle, dit modèle de Rasch[1], proposé par cet auteur dès les années 1950, on considère que les items ne peuvent se différencier que sur leur niveau de difficulté appelé paramètre **b**. Par convention on définit cette valeur de difficulté d'un item par la valeur de θ pour laquelle la probabilité de donner une réponse correcte est de P = 0,5. Ainsi dans la figure 2.4 l'item représenté a une difficulté égale à 0 (b = 0) car c'est la valeur de θ qui correspond à une probabilité de réussite de 0,5.

Les CCI des différents items sont donc toutes parallèles et ne se différencient que sur la valeur de ce paramètre *b*. Ainsi, plus la valeur de *b* augmente, plus la CCI se situe sur la droite, et plus l'item a un niveau de difficulté élevé. C'est ce qui est illustré dans la figure 2.5.

Sur la figure 2.5, les CCI des deux items sont bien parallèles mais elles sont décalées. Pour l'item *A*, qui correspond à l'item représenté sur la figure 2.4, son niveau de difficulté (paramètre *b*) est toujours de 0 ; pour l'item *B*, situé plus à droite, son niveau de difficulté est plus élevé avec un paramètre *b* ici d'environ 0,4.

➤ Le modèle à deux paramètres (indicateurs b et a)

Ce modèle de Rasch a été complexifié en 1968 par Birnbaum (cité par Dickes *et al.*, 1994) qui prend en compte les variations possibles du pouvoir

1. Mathématicien Danois.

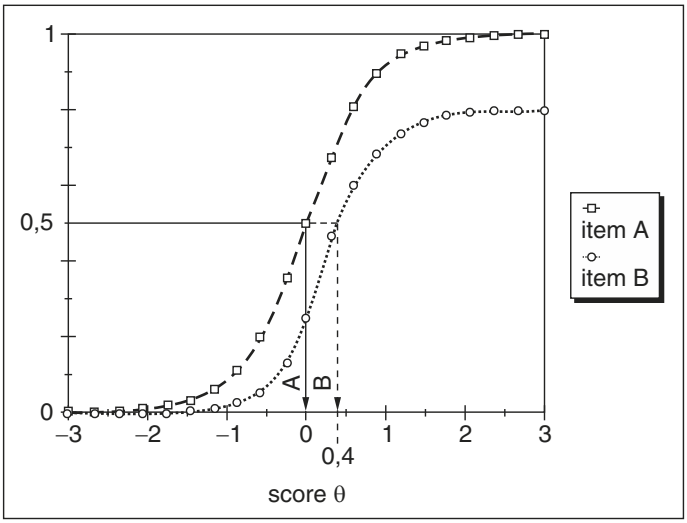

Figure 2.5
CCI de deux items de niveau de difficulté différents dans le cas du modèle MRI à un paramètre (d'après Vrignaud, 1996, p. 8, figure modifiée par nos soins).

discriminant des items. D'où un deuxième paramètre : le paramètre *a*. Il s'agit donc ici d'un modèle à deux paramètres dans lequel les items peuvent se différencier sur leur niveau de difficulté (paramètre *b*) mais également sur leur pouvoir discriminant (paramètre *a*). Cette variation dans la discrimination est représentée dans la CCI par la différence de *pente* : plus la pente est abrupte, plus l'item est discriminant. La figure 2.6 illustre ce modèle.

Sur la figure 2.6, les CCI ne sont plus parallèles car les items peuvent se différencier sur leur pente. L'item *A*, qui a la pente la plus abrupte, est plus discriminant que les deux autres items en raison d'une augmentation plus rapide des probabilités de réussite. L'item *D*, avec la pente la plus faible, est le moins discriminant des trois.

La valeur du paramètre *a* se situe habituellement entre 0 (faible discrimination) et 2 (forte discrimination). On considère qu'une valeur de 0.7 correspond à un bon niveau de discrimination (Vrignaud, 1996).

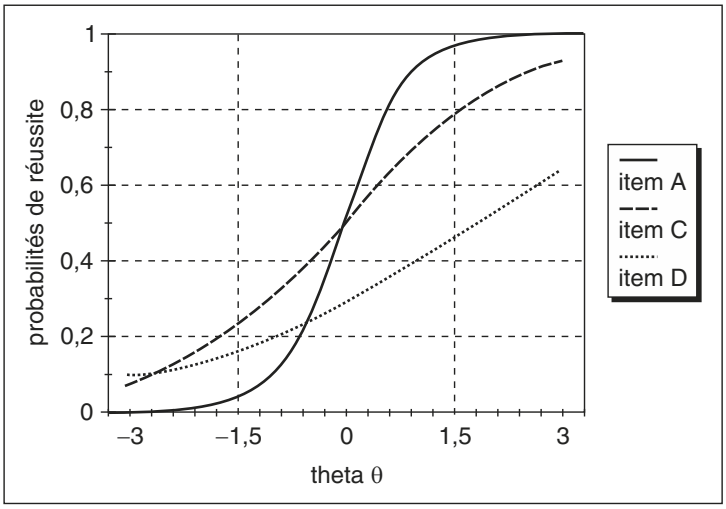

Figure 2.6
CCI de trois items dans le modèle MRI à deux paramètres (d'après Vrignaud, 1996, figure 3, p. 9).

➤ **Le modèle à trois paramètres (indicateurs b, a et c) :**

Enfin, dans certains tests, et particulièrement dans les tests de type QCM, il existe une certaine probabilité de trouver la bonne réponse en sélectionnant au hasard une réponse parmi celles proposées.

Par exemple, si dans un item donné on propose cinq possibilités de réponse, cette probabilité peut être estimée à une chance sur cinq. Le modèle a trois paramètres va donc intégrer cette nouvelle source de variation possible entre les items : le paramètre *c* ou paramètre de « pseudo-chance » (Laveault et Grégoire, 1997, p. 294).

C'est au concepteur de l'épreuve de choisir parmi ces trois modèles MRI celui qui correspond le mieux à ses objectifs et/ou aux données de l'expérimentation. Les logiciels de traitement statistique fournissent pour cela des indicateurs de conformité des données par rapport à ces trois modèles théoriques MRI. Il faudra être attentif à ces indicateurs.

Intérêts et limites des modèles MRI

Les principales limites des modèles MRI concernent leurs trois principales contraintes d'utilisation :

– nombre de sujets minimum,
– unidimensionnalité de la mesure,
– indépendance locale des items.

➤ Nombre de sujets

Pour pouvoir effectuer le calibrage des items par les modèles MRI, il est nécessaire de disposer d'un nombre important de sujets (de 600 à 1000 selon les auteurs). Ces sujets devant être de niveau de compétence variable afin d'être représentatif de l'ensemble de la population et représenter alors un ensemble assez exhaustif des différents niveaux de compétence. De plus, il convient de vérifier ensuite, sur un autre échantillon de sujets, que l'on obtient bien les mêmes indicateurs des caractéristiques des items (*a*, *b* et *c* selon le modèle choisi).

Cette exigence représente probablement l'un des freins les plus importants à l'utilisation de ces modèles.

➤ Unidimensionnalité de la mesure et indépendance locale des items

Les modèles MRI nécessitent certaines conditions mathématiques pour être utilisés, en particulier une condition d'unidimensionnalité et une condition d'indépendance locale.

L'unidimensionnalité[1] est obtenue si tous les items mesurent bien un seul et même trait : le trait latent θ (la compétence). Autrement dit, les différences individuelles entre les sujets ne doivent être expliquées que par leur différence de compétence. L'une des possibilités de vérifier cette condition est de procéder à une analyse factorielle des données qui doit aboutir à un seul facteur commun (Laveault et Grégoire, 1997).

L'indépendance locale est obtenue si la réponse à un item est indépendante de la réponse aux autres items. Cette condition implique que tous les items doivent être indépendants les uns des autres.

Par exemple, si pour donner la réponse à un item le sujet doit prendre en compte un résultat obtenu à un item précédant, cette condition n'est pas respectée. On parle alors de violation des conditions d'indépendance locale en raison des principes de construction de ces items.

1. Des modèles MRI multidimensionnels ont également été développés (Vrignaud, 1996).

Malgré ces limites contraignantes, les modèles MRI présentent plusieurs intérêts. Nous présenterons en particulier les avantages de ces modèles pour l'analyse des items, pour la possibilité qu'ils offrent de création de banques d'items, de tests sur mesure et enfin de tests adaptatifs.

➤ Analyse des items

Pour l'analyse psychométrique des items, les modèles MRI permettent différents types d'analyse : l'analyse de l'information apportée par un item, l'analyse de l'information apportée par un test et l'analyse des biais.

On considère que l'information apportée par un item est maximale lorsque son niveau de difficulté (paramètre *b*) correspond au niveau de compétence du sujet évalué : ainsi un item de difficulté moyenne (b = 0) apportera un maximum d'information pour des sujets de niveau « moyen », car un écart modéré de compétence entre ces sujets permettra de distinguer ceux qui ont une bonne probabilité de le réussir (en donnant la bonne réponse) de ceux qui ont une faible probabilité de le réussir. Tandis que pour ce même item « moyen », l'information apportée sera limitée pour les autres sujets : les sujets de faible niveau de compétence ayant une forte probabilité de l'échouer, et les sujets plus compétents une forte probabilité de le réussir.

Par ailleurs, cette quantité d'information dépend aussi du pouvoir discriminant de l'item, donc de la valeur de sa pente (paramètre *a*). Ainsi, un item peu discriminant (pente faible) avec une évolution lente de sa probabilité de réussite, nous apportera peu d'information. Par contre un item à fort pouvoir discriminant sera plus informatif. La quantité d'information apportée par chaque item peut être évaluée ainsi que le niveau pour lequel cette information est maximale (Vrignaud, 1996). Il devient alors possible de sélectionner les items les plus informatifs pour un niveau de compétence donné. À partir des estimations de l'information apportée par chaque item il est possible d'évaluer la quantité d'information du test, ainsi que le niveau de compétence où cette information est maximale.

À partir de ces analyses, il est possible de comparer différentes combinaisons d'items de façon à obtenir une épreuve correspondant à des objectifs précis (création d'épreuves sur mesure). Ainsi, par exemple, si l'objectif de l'évaluation est de sélectionner les sujets les plus performants il faudra conserver les items qui apportent un maximum d'information à un niveau élevé de compétence. Par contre, si l'objectif est d'obtenir une évaluation fine de tous les sujets, le test devra apporter de l'information sur toute l'échelle de compétence.

Enfin, concernant l'analyse des biais, les modèles MRI représentent l'une des possibilités pour repérer les items présentant un fonctionnement différentiel (F.D.I). Le principe général est le suivant : après avoir effectué l'opération de calibrage des items, on vérifie que pour deux groupes de sujets de niveau de compétence équivalent un même item ne doit pas se différencier sur ses paramètres (et en particulier sur son paramètre de difficulté). Sinon, il y a un FDI, qu'il faut alors essayer de comprendre. On trouvera dans Flieller (1999) et dans Vrignaud (2002a et 2001) des exemples d'analyse des biais par ces modèles MRI.

➤ Banques d'items, tests sur mesure et tests adaptatifs

Un autre intérêt majeur de ces modèles MRI concerne la possibilité de créer, et de gérer, des banques d'items. Qu'est-ce qu'une **banque d'items** ?

> Une banque d'items est « une collection d'items organisés, classés et catalogués, tels des livres dans une librairie, en vue de faciliter la construction d'une grande variété de tests de performance et d'autres types de tests mentaux » (Choppin, 1988, cité par Dickes *et al.*, 1994, p. 76).

À partir d'une telle « réserve » d'items, dont on connaît les principales caractéristiques (opération de calibrage des items), la construction d'un ensemble assez vaste de tests est donc possible (constitués d'items éventuellement différents, mais provenant de la même banque), adapté chacun à une population particulière et/ou un objectif déterminé. Cette possibilité de **tests sur mesure** apporte une grande flexibilité dans l'élaboration d'épreuves et/ou dans leur utilisation.

Quelques exemples :

- Il devient possible de comparer deux sujets (ou deux groupes de sujets), même s'ils n'ont pas passé les mêmes épreuves, sur leur score θ ;
- Il est également possible de créer des versions strictement parallèles de tests dont on est assuré de l'équivalence du niveau de difficulté ;
- Il est également possible de sélectionner certaines combinaisons d'items, combinaisons adaptées à des objectifs différents : c'est la notion de tests sur mesure.

L'intérêt et les limites des banques d'items ont été abordés dans l'ouvrage de Dickes *et al.* (1994, p. 76- 78).

D'une manière plus générale, ces modèles MRI apportent une flexibilité aux outils et procédures d'évaluation que ne permet pas la théorie classique des tests.

Il est également possible d'envisager des **tests adaptatifs**, le plus souvent informatisés, qui vont permettre de sélectionner les items les plus proches du niveau du sujet. Dans ce cas, le logiciel sélectionne toujours l'item le plus adapté en fonction des réponses du sujet : en cas de réussite le logiciel sélectionne alors un item plus difficile, en cas d'échec, le logiciel sélectionne un item plus facile. Ainsi, par rapport à un test classique on maximise ici le nombre d'items pertinents par rapport au niveau du sujet, et/ou on réduit le temps de passation. De plus, avec cette approche plus fine du niveau du sujet on diminue également l'erreur de mesure.

Pour des exemples concrets d'applications des MRI on peut consulter Devouche (2003) et Martin (2003). Ces deux exemples concernent des évaluations de connaissances mais le parallèle avec des évaluations psychologiques est aisé à réaliser.

Un autre intérêt de ces modèles MRI repose sur le fait que le niveau de compétence du sujet ne se définit plus comme sa position dans une certaine population (principe de l'étalonnage) mais se détermine par rapport à sa probabilité de réussir les items. D'où la possibilité d'analyser finement le contenu des items et la nature de la tâche demandée. La capacité du sujet peut donc se définir en fonction de tâches précises et non plus en référence au niveau de réussite des autres sujets. On peut alors définir plus aisément la zone de compétence du sujet, par exemple en délimitant les items pour lesquels il a une bonne probabilité de réussite (P supérieur à .70 par exemple).

Enfin, les modèles MRI permettraient de diagnostiquer de manière plus précise les sujets à haut potentiel (Caroff, 2004).

Conclusion sur les modèles MRI

Si, pour certains auteurs « ces modèles constituent un progrès majeur dans le champ de la psychométrie » (Dickes *et al.*, 1994, p. 201) d'autres auteurs, comme Reuchlin, s'interrogent sur la pertinence même de ces modèles MRI en psychologie (Reuchlin, 1997). L'une des critiques de Reuchlin concerne les bases même du modèle probabiliste. À partir de la possibilité d'évaluer le niveau des sujets par des items différents, il donne l'exemple suivant : un sujet qui fournit fréquemment des bonnes réponses à des items d'un seuil donné de difficulté sera crédité du même niveau de compétence θ

qu'un autre sujet qui fournit moins souvent des bonnes réponses à des items plus difficiles. Si, nous dit Reuchlin, il est incontestable qu'un sujet qui répond correctement, même rarement, à des questions difficiles, est capable de répondre correctement à des questions plus faciles, l'inverse n'est pas du tout évident. En effet, ce n'est pas parce qu'un sujet répond souvent correctement à des questions faciles, qu'il pourra répondre, même rarement, correctement à des questions difficiles. Reuchlin touche ici l'un des fondements des modèles MRI :

> « L'équivalence, fondamentalement postulée par le modèle, entre la fréquence des réponses et le niveau d'aptitude qu'exige chacune d'elles n'est qu'une convention formelle hautement contestable lorsqu'on passe du modèle aux réalités psychologiques auxquelles on l'applique. » (Reuchlin, 1997, p. 234).

Selon cet auteur l'application de ces modèles MRI reste limitée pour les psychologues :

> « L'étude des modèles de réponse à l'item a suscité un vif intérêt chez certains mathématiciens trouvant des thèmes de recherche dans les problèmes posés par la mesure en psychologie. Il est possible que des psychologues puissent, dans certains cas, utiliser les résultats de leurs travaux. » (Reuchlin, 1997, p. 235).

Malgré les limites indiquées, il est fort possible que dans un avenir très proche des tests utilisables en France reposent sur ces modèles MRI. Le psychologue doit alors en connaître les bases, les intérêts mais aussi les limites afin de conserver, malgré la complexification méthodologique, ses capacités d'analyse critique des outils qu'il utilise. Espérons que ces futurs tests soient accompagnés de documents explicatifs et/ou de formation had hoc favorisant cette analyse critique.

10. Conclusion

Nous avons présenté dans ce chapitre les principales notions de psychométrie utiles à tout praticien des tests, pour lui permettre une utilisation valide et raisonnée des outils qu'il utilise, notamment en ayant ce regard critique qui lui permettra de mieux en cerner les conditions d'utilisation et de mieux en maîtriser les modalités de mise en œuvre et d'interprétation.

Ce chapitre rappelle, en particulier, pourquoi il est nécessaire de respecter scrupuleusement les consignes de passation, pourquoi il est préférable (et plus fiable) d'entourer le score observé d'un intervalle de confiance, pourquoi il est utile de connaître les différents types d'étalonnages pour interpréter correctement le score d'un sujet... Il souligne également la nécessité de lire attentivement les manuels de tests et les informations qu'ils contiennent.

Ces recommandations viennent en conformité avec ce que préconise le *Code de déontologie des psychologues* (voir en annexe). Ce code insiste par exemple sur la nécessité de porter une « appréciation critique » sur les méthodes et techniques utilisés par les psychologues (article 17), et indique que les « techniques utilisées par le psychologue (...) doivent avoir été scientifiquement validées » (article 18). Nous développerons plus loin (dans le chapitre 8) les points essentiels de ce code, ainsi que ceux figurant dans les *recommandations internationales sur l'utilisation des tests*.

DEUXIÈME PARTIE

Les principaux tests d'intelligence

CHAPITRE
3

Les échelles d'intelligence

Sommaire

1. De l'échelle métrique de Binet & Simon aux échelles de Weschler

L'une des principales caractéristiques des échelles d'intelligence est qu'elles reposent sur une conception *globale* de l'intelligence et sur une approche *empirique* de sa mesure :

- *L'évaluation* porte sur les processus supérieurs (mémoire, raisonnement...) censés intervenir dans un ensemble variés de situations, ce qui explique le recours à différents types d'items pour évaluer cette capacité *globale ;*
- *La démarche* est qualifiée d'*empirique* car la sélection de ces différents types d'items repose plus sur des constats relatifs aux capacités de ces items à différencier des sujets d'âges différents que sur une approche théorique explicite de l'intelligence.

Les meilleurs représentants de cette approche, sont bien sûr Alfred Binet et Théodore Simon, ainsi que Daniel Weschler. Après une présentation synthétique de l'épreuve de Binet-Simon, nous présenterons plus en détail les échelles de Wechsler : versions WISC pour enfants et WAIS pour adultes. Toutes ces épreuves ont deux principaux points communs. Il s'agit d'une part, de leur caractère composite : ces épreuves regroupent des items très différents les uns des autres (on utilise d'ailleurs également le terme *d'échelle composite pour les décrire*), et d'autre part de la référence historique à la notion d'âge mental, qui donnera lieu ensuite à celle de Quotient Intellectuel (Q.I.).

L'échelle Métrique d'Intelligence de Binet & Simon

Nous avons évoqué dans le premier chapitre de ce livre comment, au tout début du XX^{e} siècle, suite à une demande sociale, Alfred Binet et Théodore Simon ont élaboré l'échelle métrique d'intelligence.

Pour ces auteurs, l'intelligence repose principalement sur des capacités de jugement :

> « Il y a dans l'intelligence, nous semble-t-il, un organe fondamental, celui dont le défaut ou l'altération importe le plus pour la vie pratique, c'est le jugement, autrement dit le bon sens, le sens pratique, l'initiative, la capacité de s'adapter. Bien juger, bien comprendre, bien raisonner, ce

sont les ressorts essentiels de l'intelligence. » (Binet et Simon, 1905b, p. 196-197).

Pour élaborer leur épreuve, les auteurs ont sélectionné les items selon deux critères principaux, qui illustrent les fondements de leur approche empirique :

- Ces items doivent correspondre à un ensemble varié de situations, proches de situations réelles, de façon à prendre en compte un ensemble vaste de conduites. L'ensemble de ces items forme alors une épreuve composite et hétérogène :

  « Les tests doivent être hétérogènes, cela va de soi, afin d'embrasser rapidement un vaste champ d'observation », écrivent Binet et Simon (1905b, p. 196) ;

- Ces items doivent également permettre d'identifier les enfants présentant un retard de développement mental (dénommés à l'époque enfants « anormaux ») et plus globalement, de différencier les enfants selon leur âge. Cela est rendu possible par une gradation de la difficulté des items au sein d'une épreuve progressive que l'on appelle « échelle ». La réussite à un item donné, ou à un ensemble d'items, correspond à ce que les auteurs appellent un « niveau mental ». Ce niveau mental, ou niveau de développement, correspond à ce qui sera plus tard appelé *âge mental* :

  « Bien que Binet, comme l'ont relevé Zazzo *et al.* (1966), n'utilise pas l'expression "âge mental" mais parle de "niveau mental", il peut néanmoins être considéré comme l'inventeur de la notion d'âge mental. » (Huteau et Lautrey, 1999a, p. 23).

Après une première version de l'échelle métrique d'intelligence élaborée en 1905, une seconde version paraît en 1908 puis une troisième version qui sera la version définitive en 1911. Dans cette échelle, des références en termes d'âge moyen de réussite sont indiquées, ce qui permet alors de calculer l'âge mental du sujet. Cet indicateur d'âge mental permet de situer les performances d'un enfant par rapport aux réussites moyennes observées dans les différentes catégories d'âge.

Par exemple, si un enfant âgé de 9 ans (âge biologique) réussit les épreuves qui ne sont, en moyenne, réussies que par des enfants de 10 ans, on dira alors qu'il a un âge mental de 10 ans. Il sera donc en avance d'un an dans son développement mental. À l'opposé, s'il échoue à la plupart des épreuves caractéristiques de son âge et qu'il ne réussit que les items réussis,

en moyenne, par des enfants de 8 ans, on dira alors qu'il a un âge mental de 8 ans. Dans ce cas, il présentera un retard de développement de 1 an.

Cette première échelle métrique, qui ne concernait que des enfants d'âge scolaire, va connaitre un succès considérable tant en France qu'à l'étranger, et plus particulièrement en Amérique du Nord. Une première adaptation sera réalisée aux États-Unis dès 1909 par Goddard (Huteau et Lautrey, 1999a), puis en 1916 par Terman, dans une version connue sous l'appellation « Stanford-Binet », révisée en 1937, 1960, 1972 et 1986 (Bernier et Pietrulewicz, 1997).

Cependant, en France, et malgré le succès de cette première échelle, il faudra attendre 1966 et les travaux de Zazzo et de son équipe, pour qu'une version rénovée apparaisse : la Nouvelle Échelle Métrique d'Intelligence ou NEMI (Zazzo *et al.*, 1966). Depuis cette date, aucune autre rénovation ou réétalonnage n'a eu lieu jusqu'au récent travail de Cognet et sa proposition de NEMI-II (Cognet, 2005). Cette situation explique que la NEMI ne soit actuellement quasiment plus utilisée en France, ni sans doute enseignée dans les Universités. La diffusion de la NEMI-II va peut-être redonner toute sa place à cette épreuve française.

À l'inverse de la France, paradoxalement, les adaptations du test de Binet-Simon ont bénéficié aux États-Unis de révisions régulières et de plusieurs mises à jour des étalonnages et sont encore largement utilisées, Par exemple, le Stanford-Binet, épreuve adaptée de l'échelle Binet-Simon en 1916 par Terman, psychologue à l'université de Stanford, en est actuellement à sa quatrième version (Stanford-Binet-IV). Cette dernière version de 1986, permet d'évaluer des sujets âgés de 2 ans à 23 ans à l'aide de 15 subtests[1] qui représentent quatre grandes capacités intellectuelles : raisonnement verbal, raisonnement quantitatif, raisonnement abstrait-visuel et mémoire à court terme.

Dans cette épreuve, le sujet est donc caractérisé par une note pour chacun de ces quatre domaines cognitifs, ainsi que par une note globale.

Bien que les bases théoriques de cette version Stanford-Binet-IV, et donc la fiabilité de ces quatre indicateurs, ne semblent pas être suffisamment établies (Kaufman, 2001) il est regrettable, selon nous, que les praticiens français n'aient pu bénéficier d'une telle version rénovée de l'épreuve de Binet-Simon.

Ce manque de disponibilité en France d'épreuve issue de l'échelle de Binet et Simon est sans doute l'une des raisons du grand succès des épreuves

1. En fonction de son âge, le sujet ne passe qu'une partie des subtests (de 8 à 12).

de Wechsler, qui n'a pas réellement de concurrents. L'arrivé de la NEMI-II[1] risque de modifier cet état de fait.

Les échelles de Wechsler

Daniel Wechsler (1896-1981) est avant tout un psychologue clinicien. Il est connu pour ses échelles d'intelligence qui sont parmi les tests les plus utilisés dans le monde :

– WPPSI pour les jeunes enfants ;
– WISC pour les enfants d'âge scolaire ;
– WAIS pour les adultes.

Ces différentes épreuves ont été élaborées aux États-Unis et ont ensuite été adaptées dans de nombreux pays. Nous n'aborderons ici que les versions WISC et WAIS.

L'approche de l'intelligence de Wechsler reste très proche de l'approche de Binet. En effet pour Wechsler :

> « L'intelligence est la capacité globale ou complexe de l'individu d'agir dans un but déterminé, de penser d'une manière rationnelle et d'avoir des rapports utiles avec son milieu. » (Wechsler, 1956, p. 3)

Pour évaluer cette capacité globale, il faut alors, comme dans l'échelle métrique de Binet-Simon, prendre un compte un ensemble varié de situations. C'est pour cette raison que les échelles de Wechsler comportent des situations d'évaluation assez différentes les unes des autres et constituent ainsi, comme nous l'avons déjà indiqué, des épreuves composites. Les échelles de Wechsler, comme l'échelle métrique, ont été créées afin d'évaluer une intelligence globale, une intelligence générale qui peut être définie comme la résultante d'un ensemble d'aptitudes :

> « L'intelligence générale est en effet la résultante de l'interaction d'un nombre théorique infini d'aptitudes différentes. » (Grégoire, 2000a, p. 13)

On retrouve bien ici la même volonté que celle qui était affichée par Binet de créer des épreuves hétérogènes afin de mieux évaluer cette capacité globale

1. La NEMI-II est utilisable pour des enfants âgés de 4 ans $\frac{1}{2}$ à 12 ans $\frac{1}{2}$. Elle comporte quatre épreuves obligatoires (Connaissances, Comparaisons, Matrices analogiques et Vocabulaire) et trois épreuves facultatives (Adaptation sociale, Copie de figures géométriques et Comptage de cubes).

du sujet. Nous retrouvons ici un exemple de lien étroit entre conception théorique de l'intelligence et caractéristiques du test.

Mais l'approche de Wechsler se distingue de celle de Binet sur les deux points suivants :

- L'abandon de la notion d'âge mental ;
- La distinction de deux domaines d'intelligence.

Concernant le premier point rappelons que depuis Stern, le Q.I. est un rapport (quotient) entre l'âge mental observé au test et l'âge réel (chronologique) du sujet. Ce quotient a un sens lorsque le sujet est dans une période de développement, ce qui est le cas pour des enfants. Il en a par contre beaucoup moins pour des adultes car, au-delà de 18-20 ans, le niveau de performance dans les exercices proposés ne progresse plus, en moyenne, en fonction de l'âge chronologique mais a même tendance à plafonner puis éventuellement à décliner ultérieurement sous l'effet du vieillissement (voir chapitre 1). Le calcul d'un Q.I. chez l'adulte selon la méthode de Stern, aboutirait donc à une baisse progressive du Q.I. avec l'avancée en âge. C'est l'une des raisons pour lesquelles Wechsler abandonne la notion d'âge mental et la définition du Q.I. qui en découle, et opte pour la comparaison des scores du sujet avec les scores des sujets de la même classe d'âge. Il conserve la traduction du score obtenu en Q.I. mais celui-ci n'est plus un quotient mais une indication de la place du sujet (son rang) dans une population de référence. C'est le principe de l'étalonnage qui est appliqué ici.

> Avec Weschler, le Q.I. devient un indicateur de la position du niveau du sujet par rapport à celui des sujets de même âge et non plus un indicateur relatif à une comparaison entre un âge mental et un âge chronologique.

Si Wechsler décide de conserver le terme de Q.I., c'est que cette notion est déjà très largement utilisée à cette époque. Le terme est donc une concession aux pratiques établies :

> « Le petit tour de passe-passe est d'appeler Quotient Intellectuel un score qui n'est pas un quotient mais un rang dans un groupe de référence. » Huteau et Lautrey (1999a, p. 124).

Ce nouvel indicateur Q.I se distribue selon une loi Normale (distribution Gaussienne). Pour faciliter le rapprochement avec le Q.I traditionnel, Wechsler décide de fixer la moyenne de ce nouvel indice à 100 avec un écart type de 15.

Attention !

Le Q.I. n'est pas une évaluation *absolue* du niveau intellectuel mais une évaluation *relative* de ce dernier en référence à une classe d'âge. Par exemple, si un enfant de 8 ans et un enfant de 12 ans obtiennent tous les deux la même valeur de Q.I. (par exemple, 115) il ne faut pas en conclure qu'ils ont les mêmes capacités intellectuelles ! L'enfant de 12 ans aura ici des capacités supérieures à l'enfant de 8 ans, même s'il a le même Q.I. La même valeur de Q.I., signifie simplement que ces deux enfants se situent tous les deux dans la même position vis-à-vis de leur groupe de référence, mais chacun dans son groupe d'âge. La valeur de 115 nous indique que ces deux enfants se situent au-dessus de la moyenne de leur groupe d'âge (moyenne de 100 dans chaque groupe d'âge), et plus précisément à un écart type au-dessus de celle-ci (100 + 15).

La seconde différence fondamentale entre l'échelle de Binet-Simon et les échelles de Wechsler, repose sur la présence de deux échelles distinctes, « verbale » et « performance », dans les épreuves de Wechsler car ce dernier remet en cause la forme monolithique de l'intelligence du modèle de Binet (voir chapitre 1). En effet, Wechsler considère que :

> « Les tests d'Aptitude Verbale, de Raisonnement Abstrait, et tous tests de même genre quand ils sont utilisés seuls pour un examen de l'intelligence générale, donnent seulement une image incomplète de la capacité d'un individu à s'adapter et à réussir effectivement. » (Wechsler, 1956, p. 14)

Pour pouvoir procéder à une évaluation plus complète de l'intelligence générale il décide d'intégrer d'autres types d'items et en particulier des items qui ne nécessitent pas l'usage du langage pour être résolus. Cela permet d'évaluer une intelligence non verbale (ou de réaliser une évaluation non verbale de l'intelligence). Ces items sont regroupés dans une échelle dite « échelle de Performance » alors que les autres items forment une « échelle Verbale ». Chaque échelle fait l'objet d'une évaluation séparée, avec au final, trois indicateurs des performances de sujet : un indicateur global (le Q.I.T ou Q.I. Total), et un indicateur dans chacune des échelles (Q.I « Verbal », ou QIV et Q.I de « Performance », ou QIP). Par la suite, dans les versions les plus récentes (le WISC-IV) des indicateurs factoriels viendront remplacer ces indicateurs de type Q.I.

➤ Les différentes échelles de Wechsler

La première échelle de Wechsler est diffusée aux États-Unis en 1939 sous le nom de Wechsler-Bellevue. Cette échelle d'intelligence est destinée aux adultes. Pour élaborer cette échelle, Wechsler s'inspire fortement de tests existants, et en particulier des tests de l'armée américaine *Army Alpha test* et *Army Beta Test* (Wechsler, 1956). Cette première échelle sera ensuite révisée sous le terme de WAIS (*Wechsler Adult Intelligence Scale*) en 1955, puis de WAIS-R (R pour Révisée) en 1981 et enfin de WAIS-III (3e version de la WAIS) en 1997[1].

Une version pour enfants et adolescents paraît aux États Unis en 1949 : le WISC (Wechsler Intelligence Scale for Children). Révisée en 1974 (WISC-R), puis en 1991 (WISC-III) et enfin en 2002[2] (WISC-IV).

Avant de présenter plus en détail les versions les plus récentes (WISC-III, WISC-IV et WAIS-III) interrogeons-nous sur ces rénovations : pourquoi ces échelles sont régulièrement rénovées ?

Pourquoi est-il nécessaire de rénover régulièrement les tests d'intelligence ?

Nous pouvons distinguer trois raisons principales à ces rénovations :

1° Obsolescence des items

Certains items peuvent « vieillir » au niveau du contenu, et être alors en décalage avec l'environnement actuel des sujets. Mais ils peuvent également « vieillir » au niveau de la forme : type de graphisme, type de représentation, utilisation d'images en noir et blanc...Avec un effet néfaste possible sur le niveau d'intérêt du sujet pour les épreuves, donc sur son niveau d'implication dans les tâches proposées. Par exemple, on peut remarquer l'apparition de la couleur dans certains items imagés du WISC-III alors que des représentations en noir et blanc étaient utilisées pour les items de la précédente version WISC-R.

1. Toutes les dates concernent ici les versions originales américaines. Les dates des adaptations françaises seront données plus loin dans la présentation de ces épreuves.
2. Idem note précédente.

2° Perte du pouvoir discriminant et obsolescence des étalonnages

Du fait de l'effet Flynn (voir chapitre 1), un certain nombre de tests ne permettent plus de distinguer de façon satisfaisante les sujets car ils sont devenus en moyenne trop faciles. Un relèvement du niveau de difficulté par remplacement d'un certain nombre d'items peut s'avérer nécessaire. Pour cette raison il est nécessaire d'établir régulièrement de nouveaux étalonnages. Sinon, en utilisant un étalonnage trop ancien, on risque de surestimer les capacités intellectuelles du sujet.

3° Progrès des connaissances théoriques et méthodologiques

Parallèlement à l'utilisation des tests, des études et des recherches sont menées sur les épreuves et sur les dimensions évaluées. Pour prendre en compte les résultats de ces recherches il est parfois nécessaire d'apporter des modifications concernant la structure même de l'épreuve afin de rapprocher ce qui est évalué par l'épreuve des modèles théoriques les plus récents et/ou des avancées méthodologiques. C'est pour cette raison par exemple, dont nous exposerons plus loin les éléments explicatifs, que le WISC-IV comporte de nouveaux items et de nouveaux subtests par rapport à l'ancienne version WISC-III, qui comportait lui-même de nouvelles situations d'évaluation par rapport à la version précédente WISC-R.

2. Le WISC-III

Bien que les psychologues Francophones disposent depuis 2005 de la version WISC-IV, nous avons choisi d'intégrer une présentation du WISC-III avant de présenter la version la plus récente. Il nous semble en effet nécessaire de connaître les bases du WISC-III afin de mieux comprendre les changements (importants) effectués avec l'arrivée du WISC-IV.

Le lecteur familier du WISC-III, et de l'interprétation de ses scores, pourra éventuellement survoler cette partie et/ou passer directement au chapitre suivant consacré au WISC-IV.

Le WISC[1], dans ses différentes versions, est l'échelle d'intelligence de Wechsler utilisable pour des enfants et adolescents de 6 à 16 ans. C'est

1. *Wechsler Intelligence Scale for Children.*

probablement l'un des tests les plus connus et les plus utilisés dans le monde. En France il est très utilisé, par exemple, dans le cadre de consultations en cabinet, en structure hospitalière et dans l'Éducation nationale (voir Castro *et al.*, 1996).

Nous présenterons ici les principales caractéristiques de la version WISC-III. Pour une présentation plus exhaustive on peut consulter le manuel de l'épreuve (Wechsler, 1996) ainsi que l'ouvrage très complet de Grégoire[1] qui comporte une méthodologie d'interprétation des résultats (2000a). On peut également consulter l'ouvrage de Arbisio (2003) pour une analyse des résultats à orientation plus psychanalytique.

Présentation de l'épreuve

➤ Historique

La première version WISC date de 1949, adaptée en France en 1958. Cette version a été rénovée en 1974, version WISC-R, et adaptée en France en 1981. La version WISC-III paraît ensuite en 1991, adaptée en France en 1996. La dernière version WISC-IV est sortie en 2002 aux États-Unis, avec une diffusion en France en 2005.

Ces versions sont diffusées par les ECPA.

➤ Le matériel

Le WISC-III se présente dans une mallette qui regroupe le matériel de passation. Le psychologue dispose d'un manuel très complet (294 pages) qui donne toutes les indications nécessaires à la passation et à la cotation de l'épreuve (Wechsler, 1996). Comme dans la plupart des tests à passation individuelle, c'est le psychologue qui consigne les réponses du sujet sur le cahier de passation, avec sur la première page, des espaces réservés pour reporter les scores du sujet. Le praticien peut aussi utiliser la grille d'interprétation des scores proposée par Grégoire (Grégoire, 1996).

1. Jacques Grégoire est le conseiller scientifique des ECPA pour les adaptations françaises des échelles de Wechsler.

➤ Les subtests

L'épreuve se compose de plusieurs épreuves indépendantes, appelées subtests. Chaque subtest comporte plusieurs items, présentés dans un ordre hiérarchisé en fonction de leur niveau de difficulté. Au total le WISC-III comporte 13 subtests, 6 pour l'échelle verbale et 7 pour l'échelle de performance. La passation de trois de ces subtests est optionnelle car les résultats à seulement 10 subtests sont nécessaires pour calculer les Q.I. Chaque subtest est représentatif d'un certain type de situation, d'un certain type de raisonnement.

Les subtests de l'échelle Verbale

- *Information* : il s'agit de répondre à des questions orales portant sur des connaissances que l'enfant a eu l'occasion d'acquérir. Ces connaissances sont utiles à l'enfant pour bien comprendre son environnement (30 items) ;
- *Similitudes* : la tâche consiste à trouver en quoi deux notions (ou deux objets) se ressemblent, à trouver ce qu'il y a de commun entre deux termes (19 items) ;
- *Arithmétique* : petits problèmes arithmétiques simples, à résoudre mentalement (24 items) ;
- *Vocabulaire* : consiste à indiquer la définition de mots (30 items) ;
- *Compréhension* : l'enfant doit répondre à des questions concernant des situations relatives à la vie sociale (adaptation par rapport à des situations de la vie courante) (18 items) ;
- *Mémoire immédiate des chiffres* : l'enfant doit répéter à haute voix une série de chiffres qui est lue par le psychologue (mesure de l'empan mnésique). Dans une première partie l'enfant doit répéter les chiffres dans l'ordre de présentation, dans une seconde partie il doit les répéter dans l'ordre inverse. Les séries comportent de 2 à 9 chiffres.

Les subtests de l'échelle de Performance

- *Complètement d'images* : l'enfant doit indiquer la partie manquante d'un objet représenté par une image (30 items) ;
- *Code* : l'enfant doit associer, par écrit, des chiffres à des symboles en respectant des règles d'association ;
- *Arrangement d'images* : série d'images présentées dans le désordre à remettre dans un ordre logique et chronologique (14 items) ;

- *Cubes* : reproduction de configurations géométriques à l'aide de cubes bicolores (12 items) ;
- *Assemblages d'objets* : sorte de puzzle à reconstruire (5 items) ;
- *Symboles* : l'enfant doit indiquer ici s'il retrouve des symboles cibles au milieu d'autres symboles ;
- *Labyrinthes* : l'enfant doit tracer avec un crayon un itinéraire correct dans un labyrinthe (10 items).

Standardisation

➤ La passation

La passation est individuelle et nécessite un temps d'environ 1 h 15 à 1 h 45 (durée variable en fonction de l'âge du sujet et de son niveau de réussite). La passation des subtests est effectuée selon un ordre déterminé, avec alternance entre les subtests verbaux et les subtests de performance afin de préserver chez le sujet un certain niveau de motivation. En fonction de l'âge de l'enfant il est prévu de ne pas présenter les premiers items de certains subtests (considérés comme trop faciles pour les enfants plus âgés).

Pour chaque subtest des règles d'arrêt sont aussi indiquées afin, là encore, d'adapter la passation au niveau de performance des enfants. Par exemple, dans le subtest *information* le psychologue doit arrêter la passation après 5 échecs consécutifs : on considère ainsi qu'après cinq échecs la probabilité de fournir une bonne réponse est quasi-nulle et qu'il est donc inutile de faire passer les autres items, plus difficiles. Enfin, certains subtests sont chronométrés, d'autres en temps libre. Le manuel indique très précisément les règles de passation pour chaque subtest.

➤ La cotation et les indicateurs de la performance du sujet

Cotation

Du fait du nombre de subtests, et de leur diversité, la cotation du WISC-III est plus complexe que la cotation d'un simple test de type QCM mais le manuel donne toutes les indications pertinentes pour effectuer cette cotation de manière fiable. Pour certains subtests, la cotation des items est classique : 1 point par bonne réponse et 0 point en cas d'échec. Si le temps est limité on accordera éventuellement un bonus en fonction du temps réel de réalisation,

d'où la nécessité de prendre en compte ce temps de réalisation (utilisation d'un chronomètre pour ces subtests).

Pour trois subtests de l'échelle Verbale la cotation est plus fine : on accorde 0, 1 ou 2 points en fonction de la qualité de la réponse du sujet. Par exemple, dans le subtest *similitude* si la réponse à un item repose sur « une propriété ou une fonction spécifique commune aux deux objets ou concepts » (Wechsler, 1996, p. 87) on accordera 1 point, mais si la réponse est plus élaborée et qu'elle repose sur une « généralisation pertinente relative à un aspect important des deux éléments de la paire » (Wechsler, 1996, p. 87) on accordera 2 points. De manière à garantir la fidélité de la cotation, le manuel indique, pour chaque subtest, les règles générales de cotation ainsi qu'une liste détaillée des principales réponses possibles avec les cotations afférentes.

Calcul du score par subtest

Une fois la cotation des items réalisée, on calcule les notes à chaque subtest en additionnant l'ensemble des notes obtenues aux différents items du subtest. On obtient alors un score pour chaque subtest, qui est en fait une note brute. Pour convertir ces notes brutes en notes étalonnées, dites *notes standard*, il faut consulter les tables d'étalonnage fournies dans le manuel. Bien entendu, on aura calculé au préalable l'âge du sujet afin d'utiliser les tables pertinentes. Les notes standard sont donc des notes normalisées (étalonnage normalisé), pouvant varier de 1 à 19, avec une moyenne de 10 et un écart type de 3.

Le fait que les scores à tous les subtests soient exprimés dans la même métrique (notes standard) va permettre alors d'effectuer des comparaisons du niveau de réussite du sujet en fonction des subtests (voir plus loin le détail de l'interprétation des scores).

Calcul des Q.I.

Pour le calcul des Q.I., on peut distinguer deux étapes :

– le calcul des notes de chaque échelle,
– la transformation de ces notes en Q.I.

Pour le calcul des notes d'échelles, il faut additionner, pour chaque échelle, les notes standard des cinq subtests de l'échelle. On obtient alors une note comprise entre 5 et 95, qu'il faut ensuite transformer en Q.I. grâce aux tables du manuel. On obtient alors deux Q.I., un Q.I. pour l'échelle Verbale, dit Q.I.V., et un Q.I. pour l'échelle de Performance, dit Q.I.P. Pour obtenir le Q.I. Total, ou QIT, il faut additionner les deux notes d'échelles et à nouveau consulter les tables correspondantes. Attention, ce QIT ne correspond pas

exactement à la moyenne arithmétique des deux Q.I. Par exemple, une note standard verbale de 58 (qui correspond à un QIV de 110), additionnée à une note standard de performance de 64 (qui correspond à un QIP de 120), va donner une note totale de 122, ce qui correspond à un QIT de 117 (Wechsler, 1996, Table A.4, p. 254) alors que la moyenne arithmétique des deux Q.I. est de 115.

Ces trois indicateurs de Q.I. sont exprimés dans la même métrique : moyenne de 100 et écart type de 15. Cet étalonnage normalisé des Q.I. nous permet de situer le sujet par rapport aux autres sujets de même âge : par exemple, un Q.I.T de 115 nous indique que le sujet se situe, sur cet indicateur, à un écart type au-dessus de la moyenne, ce qui signifie que ce score n'est obtenu, ou dépassé, que par 15,85 % des sujets (voir dans le chapitre 2 de ce livre les caractéristiques des étalonnages de type Q.I.).

Calcul de l'intervalle de confiance

Pour tous ces indicateurs de Q.I., comme plus globalement pour tout score à un test, il est souhaitable de prendre en compte l'erreur type de mesure afin d'entourer le score observé d'un intervalle de confiance. Rappelons que tout score observé n'est qu'une estimation de la « vraie » valeur de la compétence du sujet sur la dimension évaluée (*cf.* la théorie classique du score vrai, voir chapitre 2**)** et qu'il est préférable de caractériser le niveau d'un sujet par un intervalle de confiance plutôt que par une valeur ponctuelle correspondant au score observé. Le WISC-III, comme les autres échelles de Wechsler, est l'un des rares tests, sinon le seul, à inciter fortement le psychologue à prendre en compte cette erreur de mesure. En effet, d'une part, le manuel comporte des tableaux qui donnent pour chaque Q.I. observé (QIT mais aussi QIV et QIP) les intervalles de confiance correspondants (aux seuils .05 et .10) ; d'autre part, le psychologue doit indiquer ces intervalles en première page du protocole du sujet, juste à côté des scores observés. Par exemple, pour un QIT observé de 110, l'intervalle de confiance au seuil .10 est de 104-115. Autrement dit, si on observe un score de 110 on peut estimer que le score vrai du sujet se situera 90 fois sur 100 entre 104 et 115. Le manuel fournit donc toutes les informations permettant au psychologue de prendre en compte cette erreur de mesure.

Calcul des Indices Factoriels

Enfin, sur cette version III du WISC, il est également possible de calculer des Indices Factoriels, indices qui permettent de cerner plus précisément un aspect spécifique du fonctionnement intellectuel du sujet. Nous présenterons plus loin les bases théoriques (et les limites) de ces indices, et nous indiquons

ici uniquement les principes de calcul. Ces Indices Factoriels sont au nombre de trois :

– Indice Compréhension verbale (ou I.C.V),
– Indice Organisation perceptive (I.O.P),
– Indice Vitesse de traitement (I.V.T).

Le calcul de ces indices suit la même logique que le calcul des QIV et QIP : il faut additionner les valeurs des notes standards des subtests concernés puis consulter les tables du manuel afin de convertir ces notes en indicateurs. Ces indicateurs sont exprimés dans la même métrique que les Q.I. (moyenne de 100 et écart type de 15) et il est également prévu, dans le manuel et sur le protocole, d'entourer ces indices d'un intervalle de confiance.

Le tableau 3.1 indique le rattachement des différents items aux Q.I. et aux indices factoriels.

Tableau 3.1

Répartition des subtests[a] sur les différents indicateurs du WISC-III.

	Les 3 Q.I			Les 3 Indices Factoriels		
Subtests	QIV	QIP	QIT	ICV	IOP	IVT
Information	X		X	X		
Similitudes	X		X	X		
Arithmétique	X		X			
Vocabulaire	X		X	X		
Compréhension	X		X	X		
Mémoire des chiffres	(X)		(X)			
Complètement d'images		X	X		X	
Code		X	X			X
Arrangement d'images		X	X		X	
Cubes		X	X		X	
Assemblages d'objets		X	X		X	
Symboles		(X)	(X)			X
Labyrinthes		(X)	(X)			

a. Les parenthèses signalent les subtests optionnels.

Ce tableau nous indique que le calcul de chaque Q.I. spécifique (QIV et QIP) repose sur cinq subtests, mais que le calcul des Indices Factoriels repose sur un nombre plus faible de subtests : quatre pour ICV et IOP et seulement deux pour IVT.

De ces six scores, seul le Q.I.T prend en compte l'ensemble des subtests. C'est donc bien l'indicateur le plus complet et le plus fiable de cette échelle.

On peut remarquer également que l'Indice Factoriel ICV reprend globalement les mêmes subtests que le QIV (il manque juste le subtest *arithmétique*), de même pour l'Indice Factoriel IOP et le QIP (il manque ici uniquement le subtests *code*).

Certains subtests sont optionnels et sont destinés soit à remplacer un subtest dont le résultat n'est pas utilisable (en raison, par exemple, d'un problème lors de la passation), soit à investiguer une situation spécifique.

Enfin, on remarque également, et nous en verrons plus loin les raisons, que trois subtests (le subtest *arithmétique* et deux subtests optionnels : *mémoire des chiffres* et *labyrinthes*) ne sont rattachés à aucun Indice Factoriel.

Les étalonnages disponibles

Comme dans les autres échelles de Wechsler, on peut signaler la qualité des étalonnages : ils sont réalisés sur un nombre important de sujets, dont la représentativité est contrôlée. Pour le WISC-III, l'étalonnage Français repose sur 1 120 sujets, âgés de 6 à 16 ans. Cet étalonnage a été réalisé en 1994-1995. Ces sujets ont été sélectionnés afin de former un échantillon représentatif de la population française des enfants de cet âge (type de scolarité suivie, CSP des parents...). L'âge et le sexe ont également été contrôlés. Au final, le manuel propose des étalonnages par classes d'âges de 4 mois. Par exemple on utilisera l'étalonnage [13 ans, 4 mois, 0 jour – 13 ans, 7 mois, 30 jours] pour obtenir les notes standard d'un enfant âgé de 13 ans, 6 mois.

Les notes standard, les notes de Q.I. et les Indices Factoriels sont tous des scores étalonnés, qui suivent une loi Normale. Le tableau 3.2 indique les valeurs caractéristiques de ces indicateurs.

Tableau 3.2

Valeurs caractéristiques des indicateurs du WISC-III.

Indicateurs	Valeur mini	Valeur maxi	Moyenne	Écart type
Notes standards	1	19	10	3
Q.I.T	40	160	100	15
QIV et QIP	46	155	100	15
IOP, ICV, IVT	50	150	100	15

À partir des valeurs du tableau 3.2 il est possible de calculer très précisément la position du sujet par rapport aux sujets du même âge (voir également le chapitre 2 de ce livre). On sait par exemple que seulement environ 16 % des sujets atteignent ou dépassent la valeur seuil « moyenne + un écart type », soit 115 pour les Q.I. (et Indices Factoriels) et 13 pour les notes standard, et seulement environ 2 % des sujets atteint ou dépasse le seuil « moyenne + deux écarts types », soit 130 pour les Q.I. et 16 pour les notes standard.

Dans les tableaux d'étalonnages du manuel, les rangs percentiles sont indiqués pour toutes les valeurs des Q.I. et des Indices Factoriels. Pour les rangs percentiles des notes standard on peut consulter le tableau proposé par Grégoire (Grégoire, 2004, tableau 23, p. 163). On s'aperçoit alors, par exemple, que seulement 9 % des sujets atteignent où dépassent la note standard de 14. Il est très important que le psychologue consulte ces répartitions théoriques des scores au WISC-III afin de mieux interpréter les résultats du sujet. Par exemple, la répartition des notes standard (variation de 1 à 19) peut sembler proche de la répartition des notes scolaires (qui peuvent varier théoriquement de 0 à 20) mais en réalité, le plus souvent, cette répartition est loin d'être comparable, en raison des valeurs caractéristiques (moyenne et écart type) souvent différentes, ou inconnues, des notes scolaires. L'analogie avec les résultats scolaires est donc à éviter en l'absence de vérification de leur distribution[1].

Les qualités psychométriques du WISC-III

Le manuel du WISC-III contient de nombreuses informations relatives aux qualités psychométriques du test, assorties, le plus souvent, d'un rappel théorique des différentes notions utilisées Nous analyserons ici les informations concernant la sensibilité, la fidélité et la validité du WISC-III.

➤ La sensibilité

La sensibilité d'un test représente sa capacité à différencier les sujets. Dans un test comme le WISC-III on peut distinguer deux aspects :

1. Pour reprendre notre exemple d'une note de 14 dans un subtest du WISC-III, pour pouvoir effectuer un rapprochement avec une note scolaire de 14, il faudrait vérifier que l'on observe bien le même pourcentage de sujets (9 %) qui atteint ou dépasse cette note dans le cas d'évaluations scolaires.

- Une sensibilité *développementale*[1], qui représente la capacité du test à différencier des enfants d'âge différents ;
- Une sensibilité *interindividuelle* dans chaque classe d'âge, qui représente la capacité du test à différencier les enfants du même âge.

Au niveau de la sensibilité développementale, il faut vérifier qu'on observe bien une difficulté progressive des items de chaque subtest afin que le WISC-III puisse être véritablement considéré comme une échelle de développement. L'observation de niveaux de réussite différents en fonction de la classe d'âge permet également de déterminer, et de valider, les règles de départ et d'arrêt de chaque subtest. Règles qui permettent, rappelons-le, de ne présenter à chaque enfant que les items les plus en relation avec son niveau de compétence et de réduire au final la durée de passation. Sans rentrer ici dans le détail, on peut indiquer que les données des expérimentations présentées dans le manuel confirment ces caractéristiques développementales du WISC-III.

Au niveau de la sensibilité interindividuelle, il faut s'assurer que les scores permettent bien de différencier les enfants du même âge. Nous pouvons vérifier sur le tableau 3.2 (plus haut) que cette différenciation est assurée par l'étendue des différents scores possibles et la répartition gaussienne de ces scores. Par exemple, au niveau des notes standards, elles varient de 1 à 19 (moyenne de 10 et écart type de 3) et couvrent ainsi trois écarts types de part et d'autre de la moyenne, ce qui assure un bon niveau de différenciation des sujets. On observe cette même qualité de différenciation au niveau des Q.I. Les indices factoriels présentent une étendue un peu plus réduite que les Q.I. mais assurent un bon degré de différenciation.

➤ La fidélité

Rappelons que la fidélité concerne la constance de la mesure. Nous trouvons dans le manuel (Wechsler, 1996) des informations sur différents types de fidélité :

- La fidélité, ou consistance, *interne*, est évaluée par la méthode pair-impair. Les coefficients varient entre .64 et .84 selon les subtests, mais de .89 à .95 selon les Q.I. (p. 178). Il est tout à fait normal que les valeurs de fidélité soient plus importantes au niveau des Q.I. car ils sont constitués par davantage de scores ;

1. Dans le manuel cette sensibilité est nommée sensibilité génétique (Wechsler, 1996, p. 16).

- La fidélité *temporelle* a été évaluée par la méthode test-retest à environ 30 jours d'intervalle. Les coefficients varient ici de .57 à .89 selon les subtests, et de .87 à .94 selon les Q.I.
- La fidélité *intercotateur* varie de .95 à .99 selon les subtests, ce qui est quasiment parfait !
- L'erreur type de mesure est calculée pour chaque subtest et pour chaque indicateur global. À partir de ces valeurs, il est possible de calculer les intervalles de confiance qui entourent le score observé. Comme nous l'avons déjà indiqué, le manuel facilite ici la tâche du psychologue en donnant directement les valeurs de ces intervalles pour chaque valeur de Q.I. (aux seuils .10 et .05), ainsi que pour chaque Indice Factoriel.

Ces différents indicateurs témoignent d'un bon niveau de fidélité du WISC-III.

➤ La validité

Que mesure le WISC-III ? Quel est la fiabilité des Q.I. et celle des Indices Factoriels ? Ces questions renvoient à la validité du test.

Nous présenterons dans un premier temps des éléments d'analyse de la validité du WISC-III comme épreuve d'intelligence, puis, dans un second temps, nous questionnerons la validité de ses différents indicateurs (Q.I. et Indices Factoriels).

La validité du WISC-III comme mesure de l'intelligence

Il s'agit ici d'analyser les liaisons observées entre les scores obtenus par les mêmes sujets au WISC-III et à d'autres tests d'intelligence.

Nous trouvons dans le manuel différentes études à ce sujet, la plupart portent sur des populations américaines et sur le WISC-R, prédécesseur du WISC-III. Nous ne pouvons pas ici présenter une synthèse de toutes ces études mais nous retiendrons, d'une part, que la validité des échelles de Wechsler, et donc du WISC-III, comme mesure de l'intelligence n'est plus à démontrer (de nombreux travaux portent sur ces échelles, et confirment la validité de ces dernières), et d'autre part, que les principaux résultats des recherches portant sur le WISC-R peuvent raisonnablement être appliqués au WISC-III du fait des similitudes entre ces deux épreuves.

Sans vouloir être exhaustifs, nous ne présenterons ici que certains résultats de recherches concernant le WISC-III et principalement les données sur des populations françaises en distinguant trois approches :

- Les liaisons entre le WISC-III et les autres échelles de Wechsler (dont le WISC-R) ;
- Les liaisons avec d'autres tests d'intelligence ;
- Les liaisons entre le WISC-III et la réussite scolaire.

La première approche consiste à vérifier la nature des liaisons entre l'ancienne version de l'épreuve (WISC-R) et la nouvelle version (WISC-III). On s'attend à observer des corrélations très élevées. Le manuel du WISC-III rapporte les résultats d'une telle étude, portant sur 99 enfants.

Tableau 3.3
Validité du WISC-III : corrélations avec le WISC-R (d'après Wechsler, 1996).

	Q.I.V	Q.I.P	Q.I.T
r	.89	.80	.88

Les valeurs des corrélations observées, entre .80 et .89, témoignent bien de la similitude entre les deux versions du test : ces deux épreuves mesurent bien les mêmes domaines, aussi bien au niveau global de l'épreuve (QIT) qu'au niveau de chaque échelle (QIV et QIP).

Il est également intéressant de comparer les résultats avec les deux autres échelles de Wechsler pour des sujets situés aux extrêmes des classes d'âges. Nous trouvons là encore des données dans le manuel avec des valeurs de corrélations très satisfaisantes (voir tableau 3.4).

Tableau 3.4
Validité du WISC-III : corrélations avec la WPPSI-R et la WAIS-R (Wechsler, 1996).

	WISC-III		
	Q.I.V	Q.I.P	Q.I.T
WPPSI-R	.86	.68	.87
WAIS-R	.84	.78	.84

Enfin, dans le manuel de la dernière version de la WAIS, version WAIS-III, figurent des données avec cette version WISC-III. Là encore nous observons des corrélations importantes, au niveau des Q.I comme au niveau des deux Indices Factoriels communs aux deux épreuves. La corrélation entre les QIT est ici encore plus élevée ici avec une valeur de .91 (tableau 3.5).

Toutes ces données entre le WISC-III et les autres échelles de Wechsler confirment, s'il en était besoin, la validité du WISC-III comme mesure de l'intelligence générale.

Tableau 3.5
Validité du WISC-III : corrélations avec la WAIS-III (Wechsler, 2000).

	Q.I.V	Q.I.P	Q.I.T	I.C.V	I.O.P
r	.91	.79	.91	.88	.76

La deuxième approche consiste à comparer les résultats obtenus au WISC-III avec les résultats obtenus à d'autres épreuves d'intelligence. On s'attend ici à observer des corrélations élevées, mais inférieures à celles que nous venons de présenter : chaque test d'intelligence, s'il mesure bien le même domaine (l'intelligence) possède également ses propres spécificités (références théoriques, matériel...). Nous trouvons dans le manuel du WISC-III les résultats d'une étude comparative avec le test K-ABC. Ce dernier test permet d'évaluer deux types de processus mentaux : les Processus Séquentiels et les Processus Simultanés. La réunion de ces deux échelles permet d'évaluer un indice global d'efficience, ou Processus Mentaux Composites (PMC), qui peut être comparé au QIT du WISC. Cette épreuve comporte également une échelle de connaissance (Kaufman et Kaufman, 1993).

Tableau 3.6
Corrélations entre WISC-III et K-ABC (Wechsler, 1996).

		WISC-III					
		Q.I.V	Q.I.P	Q.I.T	I.C.V	I.O.P	I.V.T
K-ABC	P. Séquentiels	.66	.46	.64	.63	.51	.20
	P. Simultanés	.68	.61	.72	.66	.69	.22
	P.M. Composites	.74	.62	.77	.73	.69	.27
	Connaissance	.81	.46	.72	.79	.54	.19

On peut remarquer, au niveau des deux indicateurs globaux, QIT pour le WISC-III et échelle composite PMC pour le K-ABC, une valeur assez élevée de corrélation (.77) pour deux tests d'intelligence qui ne relèvent pas de la même approche théorique : malgré leurs différences, ces deux épreuves mesurent bien une même réalité.

Une autre valeur est à remarquer sur ce tableau 3.6, c'est la corrélation élevée (.81) entre QIV et l'échelle de connaissance du K-ABC. Nous y reviendrons.

Nous trouvons également, dans le manuel du WISC-III, des corrélations observées avec un autre test d'intelligence : la batterie factorielle DAT. La

batterie DAT[1] permet d'évaluer différentes aptitudes cognitives : aptitude verbale, numérique, spatiale, raisonnement... Dans l'étude présentée dans le manuel, seules trois aptitudes ont été mesurées : verbale, numérique et raisonnement abstrait.

Tableau 3.7

Validité du WISC-III : corrélations avec la DAT (d'après Wechsler, 1996).

		WISC-III		
		Q.I.V	Q.I.P	Q.I.T
DAT	Verbal	.33	.25	.31
	Numérique	.52	.47	.54
	Abstrait	.43	.37	.43

Nous pouvons observer que les corrélations sont très inférieures à celles observées dans les tableaux précédents avec des valeurs comprises entre .25 à .54 selon les indicateurs. Ce n'est pas surprenant, compte tenu du fait qu'il s'agit ici de deux épreuves très différentes : l'une, le WISC-III, évalue une intelligence générale, l'autre, la DAT, des aptitudes plus spécifiques. On pouvait cependant s'attendre à obtenir une corrélation plus élevée entre QIV et aptitude verbale (.33) qui sont, a priori, deux dimensions assez proches. Ce point est souligné dans le manuel du WISC-III :

> « La corrélation entre le Q.I. Verbal et le Raisonnement Verbal apparaît anormalement faible. Considérant le contenu des deux tests, une corrélation d'environ .50 pouvait être attendue. » (Wechsler, 1996, p. 214).

Ces résultats ne remettent nullement en cause la validité du WISC-III. D'une part, on peut effectivement remarquer, comme l'indique le manuel, que l'échantillon est restreint (l'étude porte sur seulement 41 sujets) et peut expliquer cette faiblesse. D'autre part, on pourrait également souligner que la version DAT utilisée est assez ancienne (1974 pour la version française), et qu'il conviendrait alors de mener une nouvelle étude en utilisant une version plus récente de la DAT et/ou une autre batterie factorielle.

On ne trouve, par contre, aucune étude, et on peut le regretter, qui confronte le WISC-III et un test de type facteur *g* sur une population française.

Enfin, la troisième approche de la validité consiste à comparer les résultats obtenus au WISC-III avec des indicateurs de réussite scolaire (notion

1. Une version rénovée de cette batterie, la DAT-5, a été depuis éditée (voir chapitre 5).

de validité prédictive ou concomitante). Une étude est rapportée dans le manuel qui utilise le test TNO (Test de Niveau d'Orthographe). On peut s'étonner du choix de cet indicateur de réussite scolaire qui ne prend en compte ici qu'une mesure spécifique (l'orthographe) : un test portant sur des connaissances plus larges aurait probablement mieux convenu, comme par exemple les tests TAS (Test d'Acquisition Scolaire, diffusés par les ECPA) qui permettent d'évaluer les connaissances en français mais aussi en mathématiques.

L'analyse de la liaison entre WISC-III et réussite scolaire nous paraît donc assez incomplète même si l'on observe, comme attendu, des corrélations modérées, autour de .50 (Wechsler, 1996, p. 216).

À ces données nous pourrions rajouter les liaisons observées entre l'échelle de connaissance du test K-ABC et le QIV que nous avons déjà présentées (voir tableau 3.6**).** Les valeurs élevées (.81 avec le QIV et .72 avec le QIT) témoignent également de la validité du WISC-III, et principalement ici celle du QIV, dans le domaine des acquisitions scolaires. Remarquons que ces éléments ne sont pas commentés dans cette partie du manuel du WISC-III.

La validité des indicateurs du WISC-III (Q.I. et Indices Factoriels)

Lorsqu'un test permet, comme ici, de calculer différents indicateurs du niveau de performance du sujet, il convient de vérifier les bases sur lesquelles reposent ces différents indicateurs (notion de validité structurale).

Pour les Q.I. du WISC-III, des études doivent confirmer, d'une part, la validité de la distinction de deux échelles, et donc le calcul séparé de deux scores (QIV et QIP), d'autre part, la validité d'un indice total, le QIT.

Pour les Indices Factoriels, qui n'existaient pas dans la version WISC-R, ils doivent correspondre, comme leur nom l'indique, aux facteurs mis en évidence par des méthodes statistiques appropriées (les méthodes d'analyse factorielle).

Examinons les informations fournies dans le manuel sur ces aspects.

La validité des indicateurs de type Q.I.

Le manuel propose une synthèse de différentes études qui démontrent que les regroupements des subtests pour aboutir à deux échelles distinctes, QIV et QIP, reposent sur des données solides : les liaisons sont plus importantes entre les subtests d'une même échelle qu'entre les subtests d'échelles différentes. Ce qui confirme la validité du calcul séparé des deux Q.I.

Cependant, tous les subtests présentent une corrélation significative entre eux, ce qui montre qu'ils évaluent tous une même dimension, que l'on

peut effectivement considérer comme une mesure générale de *l'intelligence*. L'addition des scores de tous les subtests pour le calcul du Q.I.T est donc aussi validée.

Nous pouvons considérer que ces trois indicateurs classiques du WISC-III (QIV, QIP et QIT) sont bien des indicateurs valides :

> « Les résultats sont relativement stables au travers des groupes et justifient clairement l'usage des QIV et des QIP au WISC-III. » (Wechsler, 1996, p. 202)

On peut remarquer à ce propos que les saturations (corrélations) des subtests sur leur échelle présentent des valeurs différentes. Le tableau 3.8 présente les subtests dans l'ordre des saturations observées.

Tableau 3.8
Saturations des subtests du WISC-III sur les QI (Wechsler, 1996).

QIV	QIP
Vocabulaire (.86)	Cubes (.74)
Information (.73)	Assemblages d'objets (.67)
Similitudes (.73)	Complètement d'images (.49)
Compréhension (.70)	Arrangements d'images (.47)
Arithmétique (.50)	Symboles (.45)
Mémoire chiffres (.33)	Labyrinthes (.41)
	Code (.39)

Les données du tableau 3.8 apportent des informations sur la force de la liaison entre chaque subtest et son échelle de rattachement (son Q.I.). Par exemple, le subtest *vocabulaire* (avec une saturation de .86) est un meilleur représentant de l'échelle Verbale que le subtest mémoire des *chiffres* (qui présente une saturation beaucoup plus faible). De même, pour l'échelle de performance, le subtest *cubes* est le meilleur représentant de cette échelle avec une valeur de saturation de .74 (voir plus loin les critères de sélection des subtests sélectionnés pour élaborer une version réduite du WISC-III).

Ces différences entre subtests se retrouvent également dans les valeurs de saturation concernant le QIT (voir Grégoire, 2000a). Ces différences, comme nous le verrons plus loin, sont à prendre en compte dans la phase d'interprétation des résultats aux subtests.

La validité des Indices Factoriels

Avant d'aborder la validité de ces indices il faut en présenter rapidement l'historique, car ces indices ne figuraient pas dans la précédente version WISC-R.

Plusieurs recherches portant sur le WISC-R mettent en évidence, par des méthodes factorielles, trois facteurs (voir par exemple les recherches de Kaufman, citées pages 193-194 dans le manuel WISC-III). Ces trois facteurs sont interprétés comme :

– la compréhension verbale, facteur qui sature la plupart des subtests de l'échelle Verbale ;
– l'organisation perceptive, qui lui sature la plupart des subtests de l'échelle Performance ;
– l'attention, ou résistance à la distraction, troisième facteur qui sature les subtests *arithmétique*, *mémoire des chiffres* et *code*.

À partir des résultats de ces recherches, il devient possible de calculer des indices représentant chacun de ces facteurs. Ces indices *factoriels* viendraient compléter les informations classiques exprimées en terme de Q.I. Mais le troisième facteur repéré semble peu fiable, car il ne repose que sur trois subtests. D'où l'un des objectifs affichés dans l'élaboration du WISC-III :

> « Renforcer la structure factorielle sous-jacente du WISC-R. » (Wechsler, 1996, p. 12).

On retrouve ici l'un des objectifs généraux, que nous avons énoncés plus haut, dans la rénovation d'épreuves : prendre en compte les résultats de recherches. Dans ce but, les concepteurs du WISC-III ont élaboré un nouveau subtest, le subtest *symboles*, qui devrait être rattaché à ce troisième facteur hypothétique, dans l'objectif d'en obtenir un indicateur plus fiable, composé alors de quatre subtests.

Les auteurs pensaient pouvoir observer ces trois facteurs dans l'expérimentation du WISC-III. Mais sur l'échantillon de sujets de la population américaine, la meilleure solution factorielle comporte quatre facteurs au lieu des trois attendus ! En effet, le nouveau subtest *symboles* s'associe en fait avec le subtest *code* pour donner au final un facteur supplémentaire. C'est pour ces raisons que la version originale (américaine) du WISC-III comporte quatre Indices Factoriels : *Compréhension Verbale*, *Organisation Perceptive*, *Attention/Concentration* (qui correspond aux subtests arithmétique et mémoire des chiffres) et ce quatrième facteur interprété comme *Vitesse*

de Traitement (qui sature les subtests *codes* et *symboles*) [voir Wechsler, 1996, p. 200-201].

Lors de l'adaptation du WISC-III à la population française, les chercheurs pensaient donc retrouver ces quatre facteurs. Mais là encore, les données sont surprenantes : s'ils retrouvent bien les facteurs Compréhension Verbale (C.V), Organisation Perceptive (O.P) et Vitesse de Traitement (V.T), le facteur Attention/Concentration ne sature plus ici que le seul subtest *mémoire des chiffres*. De plus, cette solution à quatre facteurs se révèle être instable avec l'âge. En conclusion, pour cette population française, « la fiabilité et la signification du quatrième facteur pose donc problème » (Wechsler, 1996, p. 202). C'est pour cette raison que, pour la version française du WISC-III, une structure en trois facteurs a finalement été retenue en lieu et place de la structure en quatre facteurs du WISC-III américain. C'est cette différence dans la structure factorielle des données qui explique que, comme nous l'avons déjà indiqué, trois subtests (*arithmétiques, mémoire des chiffres* et *labyrinthes*) ne sont pas utilisés dans l'adaptation Française pour le calcul des Indices Factoriels (voir tableau 3.1).

Au final, la version française du WISC-III ne comporte donc pas l'Indice Factoriel Attention/Concentration mais uniquement les trois Indices Factoriels suivants :

- Compréhension Verbale (I.C.V), qui reprend les subtests de l'échelle Verbale sauf *Arithmétique* et *Mémoire des chiffres ;*
- Organisation perceptive (I.O.P), qui reprend les subtests de l'échelle Performance sauf *Code*, *Symbole* et *Labyrinthe ;*
- Vitesse de traitement (I.V.T), formé de deux subtests *Code* et *Symbole*.

Ces trois Indices Factoriels se trouvent donc bien validés, comme nous venons de l'expliquer, par le résultat d'analyses factorielles, mais nous pouvons néanmoins émettre trois remarques les concernant.

La première concerne l'indice I.V.T. D'une part, cet indicateur ne reposant que sur deux items, sa fiabilité n'est pas assurée. D'autre part, comme l'indique Grégoire, le terme même de « vitesse de traitement » peut se discuter car ce subtest ne mesure pas exclusivement une vitesse de traitement « sa dénomination Vitesse de Traitement est sans doute trompeuse » (Grégoire, 2000a, p. 114). De plus il y a d'autres subtests dans le WISC-III qui évaluent également, au moins en partie, cet aspect du fonctionnement cognitif.

La deuxième remarque concerne la logique même de calcul de ces Indices. Nous avons vu plus haut que la liaison (la valeur de la saturation) entre un

subtest et son échelle était plus ou moins élevée en fonction du subtest (voir tableau 3.4). De la même façon, la liaison entre un subtest et son indice factoriel est plus ou moins forte. Ainsi, nous trouvons dans le manuel le classement des subtests en fonction des valeurs de saturation (Wechsler, 1996, p. 208). La logique de calcul de ces indices factoriels devrait être alors de pondérer le poids de chaque subtest en fonction des saturations observées (voir Chartier, 2001, sur ces points). Par ce mode de calcul, les indices obtenus seraient plus proches des résultats des analyses factorielles. Cette possibilité de pondération a bien été envisagée par les auteurs, mais au final, pour le calcul de ces indices, ils retiennent l'addition classique des subtests (sans effectuer de pondération) en fournissant l'explication suivante :

> « Malgré ces différences de saturations factorielles, le même poids a été attribué à tous les subtests pour le calcul des indices factoriels afin de simplifier le travail des praticiens. Ceux-ci doivent être conscients que ce mode de calcul conduit à une estimation imparfaite des facteurs. » (Wechsler, 1996, p. 208)

On ne peut que regretter cette décision car il nous semble qu'un calcul pondéré n'est pas si complexe à réaliser et permettrait au final d'obtenir des indices plus fiables.

Ces deux premières remarques concernant les limites des Indices Factoriels du WISC-III devront être prises en compte lors de la phase d'interprétation de ces indices.

Enfin, la troisième remarque concerne le problème plus général de l'adaptation des tests à d'autres populations et la recherche de biais. Nous avons abordé cette problématique dans le chapitre 2 mais nous avons ici un bel exemple de biais potentiel. Comme nous venons de le décrire, lors de la phase d'adaptation du WISC-III sur la population française, et grâce à la qualité des analyses statistiques réalisées, les auteurs ont repéré une mauvaise adéquation du modèle supposé (en quatre facteurs) et n'ont pas retenu ce modèle pour la version française de l'épreuve. Autrement dit, une adaptation moins rigoureuse de ce test aurait sans doute généralisé abusivement le calcul des quatre indices à toutes les populations. Or cette démarche serait entachée de biais. Cet exemple illustre et confirme la nécessité, lors de l'adaptation d'un test, de toujours mener des telles études.

La recherche de biais dans le WISC-III

On ne trouve guère d'études particulières sur ces aspects, sauf la référence à une étude comparative entre enfants français et enfants belges afin de vérifier que la version française du WISC-III pouvait être appliquée aux enfants belges.

Des études sur l'analyse de biais sont présentées dans l'ouvrage de Grégoire (Grégoire, 2000a). Cet auteur précise qu'elles concernent essentiellement la version WISC-R, et les populations américaines, et que globalement « *nous pouvons considérer que le WISC-R n'est pas un test biaisé pour les principaux groupes qui composent la société américaine* » (Grégoire, 2000a, p. 94). Concernant les populations francophones, Grégoire développe l'étude sur l'analyse du Fonctionnement Différentiel des Items (F.D.I) du WISC-III dans la comparaison des résultats des enfants belges et français. Nous avons présenté la notion de FDI, dans le chapitre 2. Rappelons qu'il est nécessaire de s'assurer, lors de l'élaboration d'un test, que les items sont bien du même niveau de difficulté pour des sujets de même niveau de compétence. Sinon, l'item, ou le groupe d'item, présente un FDI, et favorise alors certains sujets, ou certains groupes de sujets. Dans l'étude présentée par Grégoire, qui porte sur une version expérimentale du WISC-III, sur 40 items de l'épreuve d'information, huit ont été repérés comme biaisés (porteur de FDI). Dans la version définitive, qui ne comporte que 30 items, quatre items ont été conservés sur les huit repérés, mais dans la mesure où deux items avantagent les Français, et deux, les Belges, Grégoire considèrent que l'impact final est assez limité.

➤ Conclusion sur l'analyse des qualités psychométriques du WISC-III

Comme nous l'avons déjà indiqué, la validité du WISC-III comme mesure de l'intelligence, et plus globalement la validité des échelles de Wechsler, n'est plus à démontrer. Le manuel est complet et donne de nombreuses preuves des qualités psychométriques de ce test. Mais une lecture attentive de ce manuel nous a permis de relever quelques limites, comme par exemple celles relatives aux Indices Factoriels. Ces éléments gagneront à être pris en compte par le praticien dans la phase d'interprétation des résultats.

Les bases de l'interprétation du WISC-III

Après avoir effectué la cotation du protocole, le psychologue dispose de plusieurs indicateurs quantitatifs : les notes standards aux subtests, les trois Q.I. (QIV, QIP et QIT), sans oublier les trois Indices Factoriels (ICV, IOP et IVT). Il dispose également de données plus qualitatives qui regroupent l'ensemble des réponses données par le sujet, mais aussi ses procédures de résolution, son comportement face à une difficulté, sa motivation, son niveau de fatigabilité... Enfin, à travers les entretiens, et l'analyse éventuelle de différentes sources informations, il dispose d'autres éléments concernant le sujet (comme par exemple l'anamnèse, le dossier scolaire...).

L'étape suivante va donc consister à tenter d'articuler toutes ces informations afin de mieux comprendre le fonctionnement cognitif de l'individu singulier qui a passé le WISC-III.

Nous traiterons principalement dans cette partie de l'interprétation des indicateurs quantitatifs. Dans le cas du WISC-III, comme dans le cas des autres échelles de Wechsler, cette phase d'interprétation des résultats est relativement complexe de par la multiplicité des indicateurs et leur signification. Le praticien trouvera dans le manuel des indications assez sommaires sur le processus d'interprétation des différents scores et nous conseillons de compléter ces informations par la lecture de l'ouvrage de Grégoire qui est plus complet sur ces aspects et qui propose, à l'inverse du manuel, des études de cas (Grégoire, 2000a).

Nous présenterons ici uniquement les grandes lignes de cette phase d'interprétation car cette partie, à elle seule, mériterait un ouvrage particulier.

Le principe général d'analyse des résultats est le suivant : débuter l'analyse par l'indicateur global d'efficience, en l'occurrence ici le QIT, puis par les indicateurs plus spécifiques (QIV, QIP et les Indices Factoriels) avant d'analyser les résultats dans chaque subtest. Ce sera d'ailleurs, nous le verrons plus loin, le principe général d'analyse des autres échelles de Wechsler (WISC-IV et WAIS-III).

L'objectif général est de repérer quels sont les points forts et les points faibles du sujet examiné et, si possible, d'émettre quelques hypothèses sur les spécificités éventuelles de son fonctionnement intellectuel.

Nous pouvons d'ailleurs espérer que ce travail d'interprétation des résultats du WISC-III, à la fois riche et complexe, fasse l'objet d'une réelle formation dans le cadre des études de psychologie. Formation qui sera progressivement complétée par l'expérience acquise tout au long de l'activité professionnelle,

les stages de formation continue, les échanges avec d'autres praticiens. C'est ainsi que va se développer la compétence du psychologue dans ce domaine.

➤ Étape 1 : analyse du Q.I. Total (QIT)

La notion de Q.I.

Commençons par rappeler que le Q.I. est un indice de la position des performances[1] du sujet par rapport aux performances des sujets de même âge. Cet indicateur se distribue selon la courbe de Gauss, avec une moyenne de 100 et un écart-type de 15. Une classification des valeurs des Q.I. a été proposée dans le chapitre 2 de ce livre (tableau 2.8) . Nous retrouvons dans le manuel du WISC-III une classification analogue allant de la catégorie « retard mental » pour des Q.I. inférieurs à 69 à la catégorie « très supérieur » pour des Q.I. supérieurs à 130. Bien entendu, comme nous l'avons déjà signalé, il convient d'entourer le QIT observé d'un intervalle de confiance dans lequel va se trouver le « score vrai » du sujet.

Le tableau 3.9 reprend les catégories utilisées dans le manuel du WISC-III.

Tableau 3.9

Classification des Q.I. au WISC-III C (d'après Wechsler, 1996).

Q.I.	% théorique de sujets	Classification
130 et plus	2,2 %	Très supérieur
120-129	6,7 %	Supérieur
110-119	16,1 %	Normal fort
90-109	50 %	Moyen
80-89	16,1 %	Normal faible
70-79	6,7 %	Limite
69 et moins	2,2 %	Retard mental

Nous pouvons remarquer dans ce tableau que les catégories extrêmes sont définies par rapport à un seuil statistique :

- Moyenne plus deux écarts type[2], pour le seuil inférieur de la catégorie *très supérieure ;*

1. Performance ici est à comprendre dans son sens large et non pas en lien avec l'échelle de Performance.

2. 100 + (2x15) = 130

- Moyenne moins deux écarts types[1], pour le seuil supérieur de la catégorie *retard mental*.

Ces deux catégories ne sont donc pas définies par un type particulier de fonctionnement cognitif, mais par une proportion de population (en l'occurrence ici 2,2 % pour chaque groupe). En d'autres termes, les seuils adoptés ici, et particulièrement ceux relatifs à ces deux catégories extrêmes, même s'ils sont, comme nous le verrons plus loin, repris et utilisés dans les pratiques d'évaluation sont finalement assez arbitraires car ils ne reposent pas sur une limite identifiée entre deux types différents, ou deux niveaux distincts, de fonctionnement cognitif.

Par exemple, dans le cas de la prise en compte d'un QIT de 130 comme seuil au-delà duquel la personne sera considérée comme « surdouée », Lautrey indique bien que « ce seuil n'a aucune vertu particulière » (Lautrey, 2004, p. 227). Nous pourrions faire la même analyse à propos du seuil à partir duquel est définie la catégorie « retard mental ».

Cependant, ces seuils et les dénominations correspondantes sont, dans la pratique, largement utilisés et permettent d'interpréter de manière *qualitative* des résultats *quantitatifs*.

Le praticien dispose également dans le manuel du WISC-III de données plus précises indiquant, pour chaque valeur de Q.I. le rang percentile correspondant (Wechsler, 1996, p. 251-254). Ces informations vont permettre de situer très précisément une valeur observée de QIT.

Exemple

Prenons un sujet obtenant un Q.I.T de 112.

Une première étape consiste à entourer cette valeur d'un intervalle de confiance : à un QIT de 112 correspond l'intervalle de [106-117] au seuil .10.

Une deuxième étape consiste à situer le QIT dans la classification proposée : le sujet peut ici être catégorisé dans la classe « moyen » si on se base sur la limite inférieure de cet intervalle (106), mais il serait classé en « normal fort », si on se base cette fois sur le QIT de 112 ou sur la limite supérieure de l'intervalle de confiance (117).

Enfin, la référence au rang percentile nous permet de situer plus précisément le niveau de performance observé : à un QIT de 112, correspond le rang percentile 79, ce qui signifie que 79 % des sujets obtiennent un QIT inférieur ou égal à 112 (Wechsler, 1996, tableau A.4, p. 254).

Autrement dit, seulement 21 % des sujets du même âge obtiennent un QIT supérieur au QIT observé ici (112).

1. 100 – (2x15) = 70

Que représente le Q.I.T ?

L'interprétation de l'indicateur principal du WISC-III est à rapprocher des principes de construction de ce test et des conceptions sous-jacentes de l'intelligence de Wechsler. Ce Q.I.T est donc un indice d'une intelligence globale, d'une capacité générale d'adaptation, évaluée à travers un ensemble de tâches variées (les subtests), chacune faisant appel à un ensemble d'aptitudes diverses. Il faut alors comprendre ce Q.I.T comme étant la résultante d'un grand nombre de facteurs.

Le Q.I.T peut être considéré comme un indice proche, bien qu'un peu plus complexe, du facteur *g* de Spearman (Grégoire, 2000a).

Nous pouvons finalement retenir que cet indicateur QIT reflète le niveau global de fonctionnement intellectuel d'un individu.

Rappelons que le Q.I., et spécialement ici le Q.I.T, est souvent utilisé comme prédicteur de réussite. En effet :

> « Le QI est un des meilleurs prédicteurs de la réussite des apprentissages et des performances professionnelles. Aucune autre mesure du fonctionnement intellectuel n'a pu, à ce jour, offrir une validité prédictive supérieure. » (Grégoire, 2004, p. 83)

Nous garderons cependant à l'esprit que, même si le Q.I.T est un bon représentant de ce qui est communément appelé l'« intelligence », compte tenu du nombre limité des situations d'évaluation retenues[1] (même si elles sont variées) il ne rend pas obligatoirement compte de tous les aspects[2] de l'intelligence, notamment de ce que Wechsler appelle les « facteurs non intellectuels » de l'intelligence (voir chapitre 1) et dont il souligne l'importance (Grégoire, 2000a ; Loarer, 2006).

Q.I.T et diagnostic

Le Q.I.T est souvent utilisé, nous l'avons déjà souligné, comme critère de diagnostic pour repérer un « retard mental » ou, à l'inverse, une « précocité intellectuelle ». C'est en particulier l'épreuve la plus utilisée comme référence dans les commissions administratives destinées à orienter les élèves en grande difficulté vers les structures de l'enseignement spécialisé (voir chapitre 8 de ce livre).

Concernant *le diagnostic de retard mental* (comme déjà indiqué, *cf.* tableau 3.9), cette catégorie concerne les sujets pour lesquels un score QIT

1. Limite de tout test.
2. Ces limites sont signalées dès les premières pages du manuel du WISC-III.

maximum de 69 a été observé, ce qui représente environ 2,2 % d'une classe d'âge.

Rappelons que ce seuil est conventionnel et ne repose pas sur une définition précise de cette catégorie en termes de fonctionnement cognitif[1]. La prudence s'impose donc et le psychologue doit, avant de poser un tel diagnostic, d'une part, prendre en compte l'erreur de mesure, d'autre part, compléter cet indicateur QIT par d'autres informations concernant les capacités cognitives du sujet.

Nous rapprochons d'ailleurs cette nécessaire prudence dans le diagnostic de retard mental, avec les évolutions terminologiques dans la dénomination des scores très faibles (scores inférieurs à 69) entre la version WAIS-R qui utilisait le terme de « déficient mental » et les versions WAIS-III et WISC-IV, plus récentes, qui préfèrent utiliser le terme de « très faible ». Ce changement de dénomination, comme nous l'avons déjà évoqué dans le chapitre 2, a été adopté afin d'éviter qu'un Q.I. très faible ne soit considéré comme un indice suffisant pour établir le diagnostic de déficience mentale.

Concernant le *diagnostic de précocité intellectuelle***,** on peut remarquer que ce terme n'est pas utilisé dans la classification des résultats au WISC-III. Pour des Q.I. égaux ou supérieurs à 130, seuil qui correspond habituellement au seuil conventionnel utilisé pour repérer une « précocité intellectuelle » (Lautrey, 2004) et qui représente environ 2,2 % d'une classe d'âge[2], les auteurs du manuel du WISC-III utilisent le terme de « très supérieur » (voir tableau 3.9).

Nous devons cependant signaler que ce seuil de 130 ne fait pas l'objet d'un consensus, « selon les auteurs, ce seuil peut varier de 120 à 140, voire plus » (Caroff, 2004, p. 238). Effectivement, et nous l'avons déjà indiqué, il n'existe pas de seuil précis permettant d'identifier formellement un fonctionnement mental *qualitativement* supérieur ou différent. Ainsi :

> « Ce seuil n'a aucune vertu particulière et compte tenu de la nature conventionnelle de la définition, les discussions sur la proportion d'enfants « surdoués » dans la population ou sur la vraie valeur du QI à partir de

1. Par exemple, rien ne distingue *fondamentalement* un fonctionnement cognitif qui aboutirait à un QIT de 68, donc situé en dessous du seuil de .69/.70 d'un fonctionnement cognitif correspondant à un QIT de 72, situé lui au dessus de ce seuil.

2. C'est d'ailleurs la catégorie symétrique à la catégorie « retard mental » que nous venons de présenter.

laquelle on peut considérer qu'un enfant est « surdoué » (135 ? 150 ?) sont dénuées de sens. » (Lautrey, 2004, p. 227).

De plus, la seule valeur du QIT n'est pas suffisante car il convient d'analyser plus précisément comment ce score a été atteint. Ainsi, un QIT élevé mais obtenu avec une différence importante entre QIV et QIP, au bénéfice du QIV, n'est pas obligatoirement un indice fiable de précocité intellectuelle car ce score élevé dans l'indicateur QIV peut éventuellement résulter d'un effet de surentraînement ou de sur-stimulation du milieu, notamment du milieu familial.

Avant de porter un diagnostic de « précocité intellectuelle » il faut donc, comme dans le diagnostic de retard mental, compléter le score de QIT par la prise en compte d'autres éléments d'information sur le sujet, le QIT n'étant alors que l'un des critères disponibles pour porter un tel pronostic. Par exemple, Ziegler et Raul (cités par Caroff, 2004, p. 235) ont isolé quatre catégories de critères complémentaires au traditionnel critère de Q.I :

- les performances scolaires,
- les dimensions de la personnalité,
- la créativité,
- les intérêts.

Le lecteur intéressé par cette problématique de précocité peut consulter la revue de questions, coordonnée par Jacques Lautrey, qui a donné lieu à un numéro spécial de la revue *Psychologie Française* (Lautrey, 2004b) ainsi que l'ouvrage de Lubart (2006).

➤ Étape 2 : analyse du QIV et du QIP

L'analyse du QIT va ensuite être complétée par l'étude du profil global des résultats. On regardera en particulier si ce profil est homogène (faible différence entre QIV et QIP) ou hétérogène (différence plus importante entre ces deux Q.I.).

Le premier problème rencontré dans cette analyse est relatif aux *seuils de différences* : à partir de quelle valeur peut-on considérer que la différence entre ces deux Q.I. mérite notre attention et notre analyse ?

L'approche préconisée par le manuel est de s'appuyer sur la notion de différence significative et de proposer alors, pour chaque classe d'âge, les valeurs minimales de différences entre QIV et QIP aux seuils de signification de .15 et .05 (Wechsler, 1996, tableau B.1, p. 261).

Nous pensons, comme Grégoire (2000a), que l'on peut plus simplement considérer qu'une différence de 12 points entre QIV et QIP est suffisamment importante pour mériter une analyse. Ce seuil est à prendre avec souplesse et une différence plus faible, de 10 ou de 11 points par exemple, peut également être commentée mais avec plus de prudence dans son interprétation.

Un second problème concerne la *signification de cette différence*. Nous pouvons déjà indiquer qu'une différence de score entre QIV et QIP correspond au fonctionnement cognitif normal (non pathologique). Le manuel fournit d'ailleurs les pourcentages observés pour chaque valeur de différences et, fait qui pourrait sembler surprenant, aucun sujet de l'échantillon ne présente un QIV strictement égal au QIP (Wechsler, 1996, tableau B.2, p. 262). Au contraire, on observe une différence de 11,3 points en moyenne entre ces deux Q.I., avec des proportions non négligeables de sujets présentant des différences plus importantes. Par exemple, 16 % des sujets (soit près d'un sujet sur six) présentent une différence d'au moins 20 points entre QIV et QIP. Contrairement sans doute aux conceptions de nombre de praticiens (Grégoire, 2000a), la règle générale est bien d'observer une différence entre QIV et QIP, et non pas d'observer un profil « plat », c'est-à-dire une absence de différence entre ces deux indicateurs. Autrement dit, une différence d'une dizaine de points entre QIP et QIV est donc assez fréquente, et avant de commenter toute différence observée entre ces deux indicateurs il est conseillé de se reporter aux données du manuel afin d'estimer la singularité du profil. Ainsi ce n'est qu'à partir d'une différence QIV/QIP suffisamment importante, et suffisamment rare, que l'on pourra éventuellement envisager l'existence d'un réel trouble de type « dysfonctionnement cognitif ».

En résumé

Si le profil du sujet est homogène, avec une faible différence entre QIV et QIP, l'interprétation des résultats portera sur le QIT, les deux Q.I. spécifiques étant alors considérés comme équivalents à cet indice global.

Si le profil est plus hétérogène, avec une différence supérieure à 12 points entre QIV et QIP, il sera alors pertinent d'interpréter séparément chaque

Q.I. car le décalage observé peut refléter une réelle différence d'efficience entre les deux aspects de l'intelligence évalués par ces deux échelles.
À l'extrême, une différence trop importante entre ces deux Q.I. retirerait toute validité à l'interprétation du Q.I.T.

Que représentent le QIV et le QIP ?
Il faut ici se rappeler les bases de la construction du WISC : le QIV a été élaboré pour être une mesure de l'intelligence verbale, le QIP comme une mesure de l'intelligence non verbale (et/ou comme une mesure non verbale de l'intelligence).

On peut aussi considérer le QIV comme une estimation de l'intelligence cristallisée, le QIP étant lui plus proche de l'intelligence fluide. Même si le recouvrement entre ces différentes notions n'est pas parfait, « une équivalence approximative n'est pas absurde » (Grégoire, 2000a, p. 51).

À partir de ces informations, l'interprétation de la différence QIV/QIP dépendra, bien sûr de son importance, mais également de son sens afin de déterminer, parmi ces deux grandes dimensions de l'intelligence, quelle est celle qui semble, chez un sujet donné, plus efficiente. Bien entendu, comme pour le QI.T, l'utilisation d'un intervalle de confiance pour chaque Q.I. sera préférable à l'utilisation d'une estimation ponctuelle.

➤ Étape 3 : Analyse des résultats aux différents subtests (analyse du scatter)

La troisième étape de l'analyse consiste à repérer les résultats du sujet dans chaque subtest de manière à construire son profil de notes standard. Ce profil, sous forme de graphique, figurera d'ailleurs en première page du protocole du sujet. L'objectif général ici est de repérer les points forts et les points faibles du sujet et de tenter de les expliquer. Pour ce processus d'analyse on utilise le terme d'analyse de *scatter* (en français : analyse de la dispersion).

Comme dans l'étape précédente, l'analyse va porter sur la variabilité du profil, mais ici il ne s'agit plus d'analyser la différence entre QIV et QIP mais d'analyser les différences entre toutes les notes standards à l'intérieur de chaque échelle. En effet, un sujet peut présenter un profil de notes assez homogène, avec un niveau de réussite comparable à travers les différents subtests, ou au contraire, présenter un profil plus hétérogène avec des subtests particulièrement échoués et d'autres, au contraire, particulièrement réussis. Dans le second cas, une telle variabilité des résultats est alors souvent considérée comme l'indicateur d'un fonctionnement cognitif singulier.

Pour pouvoir interpréter ces éventuelles variations, on dispose ici de deux types de références :

- Une référence *interindividuelle* : il s'agira ici de situer le score observé à chaque subtest par rapport à la valeur moyenne de 10 (comparaison des résultats du sujet au niveau moyen de réussite observé dans la population de référence) ;
- Une référence *intra-individuelle* : il s'agira ici de situer chaque score par rapport à la moyenne individuelle du sujet (moyenne propre du sujet calculée à partir de ses différentes notes standard).

Ces deux analyses demandent à être confrontées car il faut à la fois situer le niveau de performance du sujet par rapport aux sujets de même âge, et repérer ses propres points forts et ses points faibles.

Exemple

Prenons le cas d'un sujet qui a un QIT assez élevé, avec une moyenne individuelle de 13 sur l'ensemble des subtests.

Un score de 11 à un subtest sera alors perçu comme un subtest relativement échoué (comparaison ici intra-individuelle par rapport à sa moyenne personnelle de 13) mais il ne faut pas oublier que cet « échec » est relatif car il correspond en réalité à un score plus élevé que la moyenne des sujets (comparaison ici interindividuelle par rapport à la moyenne de 10).

Dans l'interprétation du scatter il faudra toujours articuler ces deux types de comparaisons.

Pour la comparaison *interindividuelle*, il faut se rappeler que les notes standards varient de 1 à 19, avec une moyenne de 10 et un écart-type de 3. On peut donc considérer les scores supérieurs à 13 (seuil qui correspond à la moyenne + un écart-type) comme *élevés*, et les scores inférieurs à 7 (moyenne – un écart-type) comme *faibles*. Les valeurs extrêmes, correspondant à des notes déviant d'au moins deux écarts type, pouvant être alors qualifiées respectivement de *très élevées* et de *très faibles*

Le tableau 3.10 présente cette proposition de catégorisation des notes standard.

Pour la comparaison *intra-individuelle,* il est également préconisé de vérifier si la différence observée entre les subtests, ou entre les subtests et la moyenne du sujet, est suffisamment importante pour justifier une analyse. Grégoire propose ainsi la notion de « note déviante » pour définir les notes qui s'écartent significativement de la moyenne du sujet (Grégoire, 2000a).

Quelle que soit la modalité de comparaison, il convient de se rappeler que la fiabilité des interprétations d'une note isolée est limitée. Il est préférable

Tableau 3.10
Proposition de classification des notes standard au WISC-III.

Notes standard	1 à 3	4 à 6	7 à 13	14 à 16	17 à 19
Classification	Note très faible	Note faible	Note moyenne	Note élevée	Note très élevée
Répartition théorique	2,2 %	13,4 %	68,8 %	13,4 %	2,2 %

Remarque : Ce tableau ne figure pas dans le manuel du WISC-III : il s'agit d'une proposition de notre part, qui s'appuie sur des seuils statistiques et sur des propositions de Grégoire (Grégoire, 2004, p. 217).

de privilégier la prise en compte d'un ensemble plus vaste de subtests. En effet, rappelons que du fait de la conception assez empirique du WISC-III, chaque subtest ne mesure pas une et une seule aptitude, mais fait appel à un ensemble plus vaste de capacités. D'où la difficulté à interpréter de façon univoque un échec à un subtest isolé. Par contre, en analysant un ensemble de scores, en regroupant par exemple les subtests échoués, on peut alors rechercher une éventuelle cause commune qui expliquerait ces échecs.

Dans cette analyse, le praticien peut utiliser la « grille d'aide à l'interprétation des scores » qui a été proposée par Grégoire (Grégoire, 1996). Cette grille se présente sous la forme d'un tableau à double entrée avec, en colonne, les subtests du WISC-III et, en ligne, des facteurs cognitifs (aptitudes ou processus) censés intervenir dans tel ou tel subtest. Vingt-huit facteurs cognitifs sont ainsi proposés par l'auteur. Par exemple, le facteur « dépendance/indépendance à l'égard du champ » (DIC) intervient principalement, d'après cette grille, dans trois subtests : *mémoire, cubes* et *assemblages d'objets*. Un échec combiné dans ces trois subtests pourrait alors s'expliquer (ce n'est qu'une hypothèse) par l'effet de ce facteur.

Cette grille peut ainsi fournir au praticien des pistes explicatives des réussites et des échecs des sujets. Bien entendu, cette grille ne doit pas être utilisée de manière mécanique et le psychologue devra toujours formuler ses interprétations en termes d'hypothèses, qu'il conviendra de confirmer ou d'infirmer par des observations et examens complémentaires.

Même avec l'aide de cette grille, l'interprétation de la dispersion des scores (ou analyse du *scatter*) est une tâche complexe. L'explication de certains résultats peut échapper au psychologue qui doit faire preuve de prudence et d'humilité :

> « Il peut arriver que le sens de certaines dispersions de notes standard nous échappe. Il vaut alors mieux faire aveu d'ignorance plutôt que de se lancer dans des affirmations sans fondements. » Grégoire (2000a, p. 222-223).

Le psychologue doit en effet toujours étayer ses réflexions et ses interprétations sur des éléments fiables et identifiés du protocole du sujet.

D'autres approches d'analyse du *scatter* existent, la plus connue étant celles de Bourgès (1979) mais elle a été élaborée à partir de la version précédente (WISC-R). On peut également signaler les propositions plus récentes de Arbisio (2003) qui se situent plutôt dans le cadre d'une approche psychanalytique.

➤ Étape 4 : Interpréter les Indices Factoriels ?

Le lecteur attentif aura remarqué ici le point d'interrogation. En effet, pour cette version WISC-III, nous avons déjà signalé nos réserves sur l'intérêt de ces Indices Factoriels et sur les limites de leur fiabilité. En effet, comme nous l'avons déjà indiqué, l'indice ICV est très proche du QIV, l'indice IOP étant lui très proche du QIP, on peut donc raisonnablement s'interroger sur les informations spécifiques apportées par ces deux nouveaux indices.

Nous avons également apporté des éléments critiques sur les modalités de calcul de ces Indices. Enfin, nous avons déjà souligné la fiabilité insuffisante de l'indice IVT. Cette faiblesse a d'ailleurs également été relevée par Grégoire qui prône une certaine prudence dans l'interprétation de cet indice IVT (Grégoire, 2000a, p. 115).

En outre, le psychologue qui souhaiterait néanmoins interpréter ces Indices Factoriels trouvera dans le manuel du WISC-III assez peu d'information à ce sujet, ce que l'on peut regretter. Rappelons que ces indices sont exprimés dans la même métrique que les Q.I. (moyenne de 100 et écart-type de 15) et que l'on peut donc leur appliquer les mêmes principes de classification qui permettent de situer globalement le niveau de réussite du sujet (voir tableau 3.9).

Une version abrégée du WISC-III

On peut signaler qu'il existe une version abrégée du WISC-III, version qui ne figure pas dans le manuel, et qui a été proposée par Grégoire (Grégoire, 2000a). Cette version réduite a été élaborée dans l'objectif de fournir aux praticiens une épreuve plus rapide à faire passer, tout en étant suffisamment

fiable, qui puisse convenir dans certaines situations d'évaluation. Pour constituer cette épreuve, Grégoire a sélectionné les deux subtests les plus représentatifs de chaque échelle (les plus saturés avec le Q.I. de l'échelle), ce qui donne au final une version abrégée comportant seulement quatre subtests (*vocabulaire, similitude, cubes, arrangement d'images*) et ne nécessitant qu'environ trente minutes de passation. À partir de la somme de ces quatre notes standard, il est possible de calculer un indice de type Q.I. grâce aux données fournies par l'auteur (Grégoire, 2000a, p. 125).

Bien entendu, cet indicateur, basé sur seulement quatre subtests, présente une fiabilité plus faible, bien que correcte, que le Q.I.T calculé sur l'ensemble des subtests du WISC-III. C'est pour cette raison que Grégoire prône la prudence dans l'utilisation de cette forme (Grégoire, 2000a, p. 126).

Conclusion sur le WISC-III

Le WISC-III, échelle composite d'intelligence pour enfants, propose une évaluation des performances du sujet à travers des situations variées (les différents subtests), faisant ou non appel au langage (d'où la distinction entre une échelle verbale et une échelle de performance). Il s'agit d'une épreuve individuelle, basée sur une approche globale de l'intelligence, dont les indicateurs principaux sont exprimés sous la forme de trois Q.I. : QIT, QIV et QIP.

Les données du manuel concernant les qualités psychométriques du WISC-III sont nombreuses et témoignent d'une validation rigoureuse de ce test. Le praticien dispose également de trois Indices Factoriels (ICV, IOP et IVT) qui nous semblent, en l'état actuel, apporter peu d'informations supplémentaires par rapport aux indices classiques QIV et QIP.

Comme toute épreuve individuelle, le temps de passation est important, de même que le temps nécessaire à l'interprétation des résultats. En effet, une analyse fine des réussites et des échecs du sujet est possible par la méthodologie d'analyse de *scatter* (analyse de la dispersion des résultats du sujet), analyse qui vient enrichir le simple constat du niveau de performance estimé par les Q.I.

Comme nous l'avons indiqué en introduction, cette version WISC-III est remplacée depuis 2005 par le WISC-IV, version que nous allons maintenant présenter.

3. Le WISC-IV

Cette quatrième version de l'échelle de Wechsler pour enfant a été éditée en 2003 aux États-Unis et adaptée en France en 2005. Elle remplace donc la version WISC-III que nous venons de présenter.

Pourquoi une nouvelle version du WISC ?

Nous avons déjà indiqué les raisons pour lesquelles il est nécessaire de procéder régulièrement à des rénovations d'épreuve, l'une des principales étant la nécessité de disposer de normes récentes (étalonnages) afin de tenir compte de l'évolution des scores dans les tests d'intelligence (effet Flynn). Rappelons à ce propos que l'étalonnage du WISC-III datait de 1996, et qu'il était justifié alors, dix ans plus tard, de procéder à l'établissement de nouvelles normes. Mais cette réactualisation des normes ne représente que l'une des raisons de l'élaboration de cette nouvelle version WISC-IV. Plus précisément, les auteurs distinguent ici cinq objectifs principaux à cette rénovation :

- Une actualisation des fondements théoriques ;
- Une extension des applications cliniques ;
- Une meilleure adéquation développementale ;
- Une amélioration des propriétés psychométriques ;
- Une augmentation de la convivialité (WISC-IV, manuel d'interprétation, p. 8).

Mais comme nous le détaillerons dans ce chapitre, les différences apportées avec la version WISC-IV sont si nombreuses que l'on peut parler de métamorphoses, pour reprendre l'expression de Grégoire (Grégoire, 2005), voire même s'interroger sur les liaisons existantes entre cette version et l'approche originelle de Binet (Rozencwajg, 2006).

Quelles sont les principales modifications entre la version WISC-III et la version WISC-IV ?

Globalement, on peut situer ces différences à plusieurs niveaux : au niveau des subtests, au niveau de la passation, au niveau des indicateurs et enfin au niveau des règles générales d'interprétation.

➤ Modifications des subtests

Cette version WISC-IV comporte 15 subtests : 10 sont repris de l'ancienne version WISC-III et cinq sont de nouveaux subtests (le plus souvent adaptés des autres échelles de Wechsler : WPPSI-III et WAIS-III). Les subtests conservés ont fait l'objet de modifications : nouveaux items, règles d'administration et/ou de cotation, passation optionnelle...

Les 15 subtests seront présentés plus loin.

➤ Modifications au niveau de la passation

Certains subtests sont maintenant optionnels : dans cette version WISC-IV on distingue ainsi des subtests principaux et des subtests supplémentaires.

Les subtests principaux sont nécessaires pour calculer les indicateurs du test, dont le QIT, les subtests supplémentaires servant alors au calcul d'indicateurs supplémentaires, appelés « notes additionnelles » et/ou au remplacement de certains subtests obligatoires. Le praticien doit donc décider, avant ou au moment de la passation, des indicateurs qu'il souhaite recueillir afin de présenter au sujet les subtests correspondants.

➤ Modifications par rapport à la prise en compte de la vitesse de réalisation des tâches

Cette version WISC-IV contient moins de situations donnant lieu à des bonifications en fonction des temps de passation : seulement trois subtests sont concernés maintenant par cette possibilité de bonification.

➤ Modifications des indicateurs

Il s'agit là sans doute de l'une des principales modifications et en tout cas de la plus apparente apportée dans le WISC-IV : le Q.I. Total est conservé mais le Q.I. Verbal et le Q.I. Performance sont supprimés ! Les modifications touchent également les Indices factoriels : ils étaient au nombre de trois dans le WISC-III (ICV, IOP et IVT) ils sont maintenant quatre et portent des appellations différentes (voir plus loin). Le fondement de ces évolutions tient à la volonté des auteurs du WISC-IV de chercher à rapprocher ce qui est évalué par le WISC des aptitudes et processus cognitifs mis en évidence dans les modèles théoriques les plus actuels (en particulier le modèle hiérarchique proposé par Caroll que nous avons présenté chapitre 1).

Les quatre indices factoriels, nommés aussi « notes composites », évaluent chacun l'un des aspects principaux du fonctionnement cognitif du sujet :

- Indice de Compréhension Verbale (ICV),
- Indice de Raisonnement Perceptif (IRP),
- Indice de Mémoire de Travail (IMT),
- Indice de Vitesse de Traitement (IVT).

Que représentent ces nouveaux indices ? Alors que l'on peut être tenté de rapprocher ces nouveaux indices des anciens indicateurs du WISC III la vigilance est de mise : ce rapprochement est valide pour certains d'entre eux mais pas pour tous, comme nous le verrons plus loin.

➤ **Modifications des règles générales d'interprétation des résultats du sujet**

Dans le chapitre concernant le WISC-III, nous avons détaillé les règles générales d'interprétation des scores, et indiqué que celles-ci reposaient essentiellement sur l'interprétation de l'écart éventuel entre QIP et QIV. Bien entendu ici, du fait de la disparition de ces deux indicateurs, il n'est plus possible de procéder à ces comparaisons. Nous présenterons en détail les règles d'interprétation qui sont conseillées pour le WISC-IV.

Présentation de l'épreuve

➤ **Historique**

Le WISC - IV est donc la version la plus récente à ce jour de l'échelle d'intelligence de Wechsler pour enfant (*Wechsler Intelligence Scale for Children*), éditée en 2003 aux États-Unis et adaptée en 2005 en France par les ECPA. Il s'agit d'une épreuve individuelle de type échelle de développement. Elle est utilisable pour des enfants âgés de 6 à 16 ans $\frac{1}{2}$.

➤ **Le matériel**

Tout le matériel est regroupé dans une mallette (matériel de passation, de cotation, manuels...). Le psychologue dispose ici de deux manuels :

- l'un est consacré à l'ensemble des règles de passation et de cotation (WISC - IV. Manuel d'administration et de cotation, Wechsler, D., 2005a). Nous l'appellerons « manuel 1 » ;

– l'autre présente les qualités psychométriques de l'épreuve et l'interprétation des résultats (WISC-IV. Manuel d'interprétation, Wechsler, D., 2005b). Nous l'appellerons « manuel 2 »).

➤ Les subtests

Sur les 15 subtests de l'épreuve, 10 seulement sont obligatoires pour calculer les indicateurs principaux de l'épreuve (le QIT et les quatre notes composites). Cinq subtests sont ainsi optionnels (notes additionnelles), destinés à fournir des informations supplémentaires (en fonction des objectifs du psychologue), et éventuellement à mesurer un indice spécifique dans le cas du subtest Barrage. Ces subtests optionnels peuvent, dans certains cas, remplacer des subtests obligatoires dont les résultats ne seraient pas valides.

Les auteurs du manuel conseillent ainsi de faire passer systématiquement les subtests Arithmétiques et Barrages (manuel 1, p. 27), ce qui fait au final 12 subtests à faire passer.

Nous allons présenter les subtests à partir de leur indice de rattachement (note composite).

Les subtests de l'Indice de Compréhension Verbale (ICV)

- *Similitudes* : directement issu du WISC - III ce subtest consiste à trouver en quoi deux notions (ou deux objets) se ressemblent (chercher les similitudes) (23 items, dont 15 nouveaux) ;
- *Vocabulaire* : subtest classique du WISC dans lequel l'enfant doit indiquer la définition de mots (36 items, dont 27 nouveaux) ;
- *Compréhension* : issu lui aussi du WISC - III ce subtest évalue la capacité de l'enfant à expliquer des situations de la vie courante (21 items dont 13 nouveaux) ;
- *Information* : subtest obligatoire dans le WISC - III il devient ici l'un des subtests optionnels de l'échelle. Dans ce subtest l'enfant doit répondre à des questions de connaissances (33 items, dont 19 nouveaux) ;
- *Raisonnement verbal* : second subtest supplémentaire c'est également un nouveau subtest dans lequel il s'agit de répondre à des devinettes (24 items).

L'indice ICV est donc constitué de trois subtests principaux et de deux subtests supplémentaires, dont l'un est entièrement nouveau (Raisonnement verbal).

Les subtests de l'Indice de Raisonnement Perceptif (IRP)

- *Cubes* : subtest classique des échelles de Wechsler dans lequel l'enfant doit reproduire une configuration géométrique à l'aide de cubes bicolores (14 items, dont 3 nouveaux) ;
- *Identification de concepts* : nouveau subtest dans lequel l'enfant doit choisir une image afin de constituer un groupement autour d'un concept commun (28 items) ;
- *Matrices* : nouveau subtest de type matrice analogique, adapté de la WAIS-III, et proche des tests de type facteur *g* (35 items) ;
- *Complètement d'images* : ce subtest qui était obligatoire dans le WISC-III devient ici supplémentaire. L'enfant doit indiquer, ou nommer, la partie manquante d'un objet représenté par une image (38 items, dont 13 nouveaux).

L'indice IRP est donc constitué de trois subtests principaux et d'un subtest supplémentaire.

Les subtests de l'Indice de Mémoire de Travail (IMT)

- *Mémoire de chiffres* : subtest issu du WISC-III, dont la tâche consiste ici à répéter une suite de chiffres. On peut remarquer que ce subtest, optionnel dans le WISC-III, devient ici obligatoire. Ce subtest est organisé en deux séries : ordre direct et ordre inverse (8 items dans chaque série) ;
- *Séquence Lettres-Chiffres* : nouveau subtest, composé comme son nom l'indique de lettres et de chiffres, et adapté de la WAIS-III, dans lequel le psychologue lit une séquence au sujet qui doit ensuite restituer les chiffres, puis les lettres (10 items) ;
- *Arithmétiques* : subtest obligatoire du WISC-III, il devient ici supplémentaire. Il s'agit de traiter mentalement des petits problèmes arithmétiques (34 items dont 24 nouveaux).

L'indice IMT est donc constitué de deux subtests principaux et d'un subtest supplémentaire.

Les subtests de l'Indice de Vitesse de Traitement (IVT)

- *Code* : subtest issu du WISC-III dans lequel l'enfant doit copier des symboles associés à des figures géométriques. En fonction de l'âge du sujet deux formes existent : code A et code B ;
- *Symboles* : également issu du WISC-III, mais avec ici un statut de subtest obligatoire, la tâche consiste à indiquer si un symbole proposé correspond à l'un des symboles cibles. Deux formes également ici en fonction de l'âge de l'enfant : symbole A et symbole B ;

- *Barrage* : subtest nouveau et optionnel dans lequel l'enfant doit barrer des images cibles.

L'indice IVT est donc constitué de deux subtests principaux et d'un subtest supplémentaire.

Par rapport au WISC-III, les changements relatifs aux subtests sont donc très nombreux :

1. Apparition de nouveaux subtests ;
2. Disparition de certains subtests (en particulier *Arrangements d'images* et *Assemblages d'objets)* ce qui diminue fortement le nombre de subtests qui reposent sur une manipulation concrète de matériel (il ne reste que le subtest cubes) ;
3. Ajout ou changement d'items aux anciens subtests (parfois en proportion très importante) ;
4. Changement de statut (obligatoire ou optionnel) de certains subtests ;
5. Regroupement des subtests en quatre indices (ou notes composites), et disparition des échelles verbale et performance.

Ces modifications ne sont pas sans conséquences sur ce qui est évalué par le WISC-IV (Grégoire, 2005 ; Rozencwajg, 2006).

Standardisation

➤ La passation

La passation est individuelle et nécessite un temps de passation compris entre 1h15 et 1h45 en fonction du niveau de réussite de l'enfant.

La durée de passation peut également varier, et ceci est nouveau, en fonction des décisions du psychologue. En effet, un certain nombre de subtests sont optionnels et leur passation va dépendre des objectifs du psychologue. Si celui-ci souhaite calculer uniquement les quatre indices centraux du WISC-IV, la passation des 10 items principaux est suffisante ; par contre, s'il souhaite pouvoir calculer également une, ou des, note(s) additionnelle(s) alors la passation doit comprendre les subtests supplémentaires correspondants.

L'ordre de passation des subtests est codifié avec une alternance entre différents types de tâches.

La passation de chaque subtest débute par des items-exemples, et se poursuit en fonction de l'âge des sujets (règles de départ spécifiques à chaque subtest). Toutes les précisions concernant les règles de passation figurent dans le manuel 1 (pages 65 à 203).

➤ La cotation : les indicateurs de la performance du sujet

Les règles générales de cotation du WISC-III sont reprises :

– cotation en 0 ou 1 pour certains subtests ;
– cotation plus fine en 0, 1 ou 2 pour d'autres ;
– bonification éventuelle selon la vitesse d'exécution...

De même est conservé le processus d'élaboration des notes standard (notes étalonnées en référence aux enfants de même âge) à partir des scores bruts du sujet. Rappelons que les notes standards de chaque subtest peuvent varier de 1 à 19, avec une moyenne de 10 et un écart-type de 3. Cette standardisation des notes rend directement possible les comparaisons du niveau de performance du même sujet sur des subtests différents (variations intra-individuelles et analyse de *scatter*).

Le regroupement des subtests par échelle permet le calcul des quatre indicateurs ICV, IRP, IMT et IVT, puis de l'indicateur global QIT. Comme dans le WISC-III tous ces indicateurs adoptent la même métrique : moyenne de 100 et écart-type de 15.

Au final on dispose donc d'une note (standard) pour chaque subtest et d'un score pour chaque indice.

Le tableau 3.11 permet de synthétiser la structure du WISC-IV.

Comme le montre ce tableau, le calcul de chaque note composite repose sur un nombre limité de subtests (deux ou trois) alors que l'indicateur global, le QIT, prend en compte l'ensemble des 10 subtests obligatoires. Cet indice est donc bien l'indicateur le plus complet, et le plus fiable, de l'épreuve WISC-IV.

Les nouveaux indicateurs « notes additionnelles » s'expriment dans la même métrique que les notes standards et permettent d'obtenir des informations plus précises sur certaines aptitudes cognitives (voir plus loin).

Tableau 3.11
Répartition des subtests sur les différents indicateurs du WISC–IV.

	Les indicateurs du WISC-IV					
	Les notes composites				L'indicateur global : QI T	Les notes additionnelles
Subtests	ICV	IRP	IMT	IVT		
Similitudes	X				X	
Vocabulaire	X				X	
Compréhension	X				X	
Information	(X)				(X)	
Raisonnement verbal	(X)				(X)	
Cubes		X			X	sans bonification
Identification de Concepts		X			X	
Matrice		X			X	
Complètement d'images		(X)			(X)	
Mémoire de chiffres			X		X	ordre direct
						ordre inverse
Séquence Lettres-Chiffres			X		X	
Arithmétique			(X)		(X)	
Code				X	X	
Symboles				X	X	
Barrage				(X)	(X)	ordre aléatoire
						ordre structuré

(Les parenthèses signalent les subtests optionnels.)

Les étalonnages disponibles

L'étalonnage repose sur un échantillon de 1 103 enfants, âgés de 6 à 16 ans, représentatifs de la population française. Comme pour les autres échelles de Wechsler, il faut souligner ici l'attention apportée à la constitution de l'échantillonnage de sujets avec contrôle de différentes variables : profession et CSP des parents, zone d'habitation, sexe et âge des enfants... Au final on dispose d'étalonnages par classes d'âges, avec des groupes d'âges de 4 mois (voir annexe A du manuel 1).

Comme pour les autres échelles de Wechsler les indicateurs du niveau de performance du sujet sont des scores étalonnés, avec une moyenne de 10 pour les notes standards de chaque subtest et une moyenne de 100 pour les indicateurs principaux de l'échelle. Le tableau 3.12 donne les valeurs caractéristiques de chaque indice.

Tableau 3.12

Valeurs caractéristiques des indicateurs du WISC-IV.

Indicateurs	Valeur mini	Valeur maxi	Moyenne	Écart type
Notes standards et notes additionnelles	1	19	10	3
Q.I.T	40	160	100	15
Notes composites : ICV, IRP, IMT, IVT	50	150	100	15

Le manuel 1 fournit également les rangs percentiles et les intervalles de confiance, pour les indicateurs factoriels et pour le QIT. Il est intéressant de consulter ces tables afin de situer plus précisément les performances du sujet : par exemple, il faut savoir qu'un QIT de 110 n'est atteint (ou dépassé) que par seulement 30 % des enfants (voir également le chapitre 2 ce livre sur les étalonnages de type Q.I.).

Les qualités psychométriques

Nous reprendrons ici les indications du manuel concernant les qualités psychométriques du WISC-IV (sensibilité, fidélité et validité) en effectuant des comparaisons avec l'ancienne version WISC-III.

Rappelons que la version originale, éditée aux États-Unis, a fait l'objet d'une validation et qu'il ne s'agit donc ici que de vérifier les qualités psychométriques de l'adaptation française sur l'échantillon d'enfants français.

➤ La sensibilité

Il faut ici distinguer deux aspects :

- La sensibilité au sens classique du terme, c'est-à-dire la capacité du WISC-IV à différencier des enfants du même niveau d'âge ;

- La sensibilité développementale : les échelles de Wechsler étant des échelles de développement, il faut ici vérifier que le niveau moyen de réussite des items est bien ordonné selon l'âge moyen des sujets et permet donc de différencier des enfants d'âge différents.

Pour le premier aspect de la sensibilité, les données du tableau 3.12 indiquent que cette sensibilité est assurée au niveau de tous les indices. Remarquons que l'indice QIT, avec des valeurs possibles de 40 à 160, permet, par rapport aux notes composites (ICV, IRP, IMT et IVT), une différenciation plus fine des enfants situés dans les catégories extrêmes. Les notes standards, comme les notes composites, présentent néanmoins un bon niveau de sensibilité avec des valeurs s'échelonnant sur trois écarts types de part et d'autre de la moyenne.

➤ La fidélité

La fidélité, ou consistance interne a été évaluée à partir de la méthode pair-impair. Les coefficients varient de .65 à .86 pour les subtests, de .62 à .82 pour les notes additionnelles et de .84 à .89 pour les notes composites. Cette fidélité est de .94 pour le QIT.

La fidélité temporelle a été évaluée par la méthode test-retest sur un échantillon de 93 enfants avec un intervalle moyen de 27 jours entre les deux passations. Les valeurs sont globalement correctes avec des variations de .64 à .83 selon les subtests, de .78 à .88 selon les notes composites et une valeur de .91 sur le QIT (manuel 1, p. 34). On observe également, et c'est attendu, des gains moyens entre les deux passations (effets d'apprentissage).

La fidélité de la cotation sur les subtests verbaux a également fait l'objet d'une évaluation, avec des valeurs quasiment parfaites : .98 et .99.

Enfin, l'erreur de mesure, inversement proportionnelle à la fidélité du test, a été évaluée pour chaque type d'indicateur. Exprimée en unité d'écart-type, cette erreur-type de mesure varie pour les notes standards de 1,16 à 1,78, de 4,98 à 6,01 pour les notes composites et elle est estimée à 3,63 pour le QIT. À partir de ces indications, il devient possible de calculer un intervalle de confiance dans lequel doit se situer la note vraie du sujet. Comme pour le WISC-III, les auteurs du manuel nous facilitent la tâche en proposant les valeurs de cet intervalle, pour les risques de 5 % et de 10 %, et pour chaque indice (en annexe du manuel 1). Le praticien est d'ailleurs invité à faire figurer pour chaque score de l'enfant un intervalle de confiance.

L'ensemble des données concernant la fidélité du WISC-IV témoigne d'un bon niveau de fidélité de l'épreuve mais le praticien doit se rappeler que le QIT est l'indicateur qui présente le meilleur niveau de fidélité et que les scores aux indices composites sont plus fidèles que les scores aux subtests.

➤ La validité

Rappelons que c'est sans doute la qualité la plus importante d'une épreuve. Il s'agit ici d'analyser les données qui justifient l'utilisation du WISC-IV comme mesure de l'aptitude intellectuelle. Dans un premier temps nous analyserons les données concernant l'analyse de la validité du WISC-IV comme mesure de l'intelligence puis, dans un second temps, nous nous intéresserons à la validité de la structure de l'épreuve (la validité des différents indicateurs du WISC-IV).

La validité du WISC-IV comme mesure de l'intelligence

Bien que la validité des échelles de Wechsler soit attestée par un grand nombre d'études publiées, il est normal que, lors de chaque rénovation d'épreuve, les auteurs apportent des éléments de validation concernant la nouvelle version. Ce sont ces éléments que nous allons présenter et analyser. Comme nous l'avons déjà indiqué (voir le chapitre 2 de ce livre), la validation est à entendre comme un processus : les éléments présents dans le manuel vont être progressivement complétés par les publications d'études sur cet instrument Nous nous centrerons ici principalement sur les données concernant l'échantillon français. Une première approche consiste à comparer le WISC-IV avec les autres échelles de Wechsler, une seconde approche consistera à analyser les liaisons existant avec d'autres tests d'intelligence.

Corrélations avec le WISC-III

On attend des corrélations élevées entre les deux versions de l'épreuve même si, comme nous l'avons déjà signalé, les différences sont nombreuses entre WISC-III et WISC-IV. Ces deux tests ont été administrés à 159 enfants. La corrélation obtenue sur les QIT est de .78, valeur élevée mais cependant un plus faible que la corrélation qui avait été observée entre WISC-III et WISC-R, qui était de .88 (voir tableau 3.3). Cette baisse de corrélation peut s'expliquer par les modifications importantes apportées au WISC-IV.

Malgré cela, la valeur élevée de la corrélation indique que ces deux épreuves évaluent bien le même domaine : l'intelligence dite globale.

Il est également intéressant d'observer les corrélations entre les différents indicateurs de ces deux versions. C'est ce que nous permet le tableau 3.13.

Tableau 3.13

Corrélations entre WISC-IV et WISC-III (d'après le manuel 2).

		WISC-IV				
		QI T	ICV	IRP	IMT	IVT
WISC-III	QI T	.78				
	QI V		.82			
	QI P			.62		
	ICV		.83			
	IOP			.60		
	IVT					.66

Que nous apportent ces valeurs ? Elles permettent d'estimer les relations qui existent entre les indicateurs de la version WISC-III et les nouveaux indicateurs (les notes composites) de la version WISC-IV. Par exemple, l'indicateur ICV du WISC-IV est, comme attendu, assez proche à la fois du QIV (r = .82) et de l'ICV (r = .83) du WISC-III. Par contre les liaisons entre le nouvel indicateur IRP du WISC-IV et les indicateurs les plus proches du WISC-III sont moins élevées : .62 avec le QIP et .60 avec l'indice IOP.

Ces observations confirment ici un point que nous avons déjà évoqué : il ne faut pas chercher à assimiler les indicateurs des deux versions et principalement, on le voit ici, l'indice IRP au QIP. En effet, et c'est un point que nous reprendrons, même s'il existe un assez large recouvrement entre les deux indices, ce qui est évalué par l'indice IRP du WISC-IV ne correspond qu'en partie à ce qui était évalué par l'indice QIP du WISC-III.

Corrélations avec la WPPSI-III

Rappelons que la WPPSI-III est l'échelle d'intelligence de Wechsler destinée aux enfants plus jeunes, âgés de moins de 7 ans. Ces deux épreuves ont été administrées à 60 enfants âgés de 6 à 7 ans. Les résultats figurent dans le tableau 3.14.

Les corrélations observées ici (de .69 à .84) sont globalement du même niveau que celles qui avaient été observées entre le WISC-III et la WPPSI-R (voir tableau 3.4) et témoignent de la proximité de ce qui est évalué par les nouvelles versions de ces deux instruments.

Tableau 3.14
Corrélations entre WISC-IV et WPPSI-III (d'après le manuel 2).

		WISC-IV				
		QI T	ICV	IRP	IMT	IVT
WPPSI-III	QI T	.84				
	QI V		.79			
	QI P			.72		
	QVT					.69

Corrélations avec la WAIS-III
Il s'agit là encore de comparer les résultats entre deux versions des échelles de Wechsler mais cette fois pour les sujets les plus âgés. L'échantillon est composé ici de 55 sujets âgés de 16 ans. La corrélation de .83 observée entre les QIT (tableau 3.15) confirme également que le WISC-IV évalue bien la même forme d'intelligence que celle évaluée par la WAIS-III.

Tableau 3.15
Corrélations entre WISC-IV et WAIS-III (d'après le manuel 2).

		WISC-IV				
		QI T	ICV	IRP	IMT	IVT
WAIS-III	QI T	.83				
	QI V		.81			
	QI P			.74		
	ICV		.78			
	IOP			.78		
	IMT				.79	
	IVT					.64

Nous avons déjà indiqué que la nouvelle structure du WISC-IV, avec ses quatre notes composites, est très proche de la version WAIS-III qui contient également quatre indices composites. La proximité de ces indices (pris deux à deux) confirment l'équivalence structurelle des deux épreuves avec des corrélations élevées entre les mêmes indicateurs : .78 au niveau de l'indicateur de la compréhension verbale (ICV), .78 également au niveau de l'organisation/raisonnement, perceptif (IRP/IOP), .79 au niveau de la mémoire de travail (IMT) et .64 pour la vitesse de traitement (IVT).

Corrélations avec le K-ABC

Les deux épreuves ont été administrées à 70 enfants âgés de 6 à 12 ans. Rappelons que le K- ABC, qui relève d'une autre approche que celle de Wechsler, distingue deux types de processus cognitifs :

– les processus séquentiels (P. Séquentiels),
– les processus simultanés (P. Simultanés).

L'échelle des processus mentaux composites (PMC) représente l'indicateur global de cette épreuve, qui comporte également une échelle de connaissance (Kaufman et Kaufman, 1993).

Tableau 3.16

Corrélations entre WISC-IV et K-ABC (d'après le manuel 2).

		WISC-IV				
		Q.I.T	ICV	IRP	IMT	IVT
K-ABC	P. Séquentiels	.59	.36	.49	.70	.30
	P. Simultanés	.57	.38	.64	.33	.41
	P.M. Composites	.69	.44	.70	.58	.44
	Connaissances	.74	.66	.69	.61	.32

Au niveau des deux indicateurs globaux, QIT pour le WISC-IV et échelle composite PMC pour le K-ABC, la corrélation est de .69. Cette valeur témoigne du large recouvrement de ce que mesurent ces deux tests, bien qu'ils reposent sur des approches théoriques différentes.

On aurait pu s'attendre à observer une valeur plus élevée entre K-ABC et WISC-IV qu'entre K-ABC et WISC-III compte tenu que le WISC-IV affiche clairement un ancrage cognitif. Or c'est l'inverse qui est observé (même si la différence reste faible) : la corrélation entre K-ABC et WISC-III était, sur ces mêmes indicateurs, de .74 (voir tableau 3.6).

On peut également remarquer la valeur élevée (.74) de la corrélation entre QIT et l'échelle de connaissance du K-ABC, valeur plus élevée que celle observée entre les deux indicateurs d'intelligence (.69 entre QIT et P.M.C).

Enfin, on observe une proximité plus importante, d'une part entre IRP et Processus Simultanés (.64, contre une valeur de .49 avec Processus Séquentiels), d'autre part, entre IMT et Processus Séquentiels (.70, contre .33 avec Processus Simultanés). Les autres indices, QIT, ICV et IVT, étant plus équilibrés à ce niveau.

Toutes les études de validité présentées dans le manuel (avec le WISC-III, la WPPSI-III, la WAIS-III et le K-ABC) apportent donc des éléments

convergents sur la fiabilité du WISC-IV comme mesure de l'intelligence globale.

Par contre ne figurent ici aucune étude concernant les liaisons entre WISC-IV et batteries factorielles[1], ni entre WISC-IV et épreuve de facteur *g*. On ne peut que le regretter.

Concernant cette fois la validité critérielle, on s'attend à observer des données concernant par exemple les liaisons entre WISC-IV et réussite scolaire. Étonnamment aucune étude de ce type, sur une population française, n'est présentée dans le manuel. Rappelons, là aussi, que de telles études avaient été menées lors de l'adaptation du WISC-III, avec l'analyse des relations entre les indicateurs du WISC et des tests standardisés de connaissance. Sachant que le WISC est très utilisé dans le cadre scolaire il est fort regrettable que de telles études ne figurent pas dans le manuel de cette version WISC-IV. Même s'il est fort probable que l'on puisse appliquer au WISC-IV les résultats observés avec le WISC-III, la présentation de données permettrait de conforter, et de préciser, la validité du WISC-IV dans ce domaine.

Les seules données disponibles sont celles observées entre WISC-IV et l'échelle de connaissance du K-ABC, données que nous venons de présenter (voir tableau 3.16) et qui témoignent de la relation étroite entre ce qui est évalué par le WISC-IV et les connaissances acquises (corrélation de .74 entre QIT et l'échelle de connaissance du K-ABC). Il est d'ailleurs un peu surprenant de constater sur ce tableau 3.16 que, contrairement à ce que l'on pourrait prédire, entre ICV et IRP, c'est l'indicateur théoriquement le moins lié aux connaissances (IRP) qui présente la corrélation la plus élevée avec cette échelle de Connaissances du K-ABC (.69 contre .66). Même si la différence reste faible, ce résultat mériterait une analyse, ou au moins un commentaire, tous deux absents du manuel.

Études de groupes cliniques

Cette partie constitue également une nouveauté par rapport à l'ancienne version WISC-III. En effet, même si nous disposions de données et/ou de pistes concernant les possibilités d'utilisation du WISC-III dans une démarche de premier diagnostic de différents troubles (voir Grégoire, 2000a) il faut signaler la présence ici d'études spécifiques sur différents types de sujets : précocité intellectuelle, épilepsie, dyslexie, retard mental léger,

1. Rappelons qu'une telle étude, avec la batterie DAT, a été menée dans l'expérimentation du WISC-III (voir tableau 3.7).

troubles des apprentissages, trouble de l'attention/hyperactivité, trouble du langage, traumatisme cérébral, trouble autistique...

Ces études concernent majoritairement des données américaines (seules trois études[1] sur 18 portent sur des enfants français), le plus souvent sur de petits effectifs, avec des critères de diagnostic pouvant être différents. Par conséquent, il faut prendre avec prudence les observations et conclusions de ces études, comme d'ailleurs le préconisent les auteurs du manuel (manuel 2, p. 63). Nous indiquerons ici uniquement le principe général de ces études : il s'agit de comparer les résultats d'un groupe clinique à un groupe témoin et de relever les éventuelles spécificités du profil de ce groupe.

> ***Exemple***
>
> Le groupe d'enfants dyslexiques se singularise par des notes standards particulièrement faibles dans les subtests reposant sur le langage, comme *information* (m = 5,9) et *vocabulaire* (m = 6), mais également sur des subtests qui font le plus appel à la mémoire de travail : *arithmétiques* (m = 5,2) et *Séquence Lettres-Chiffres* (m = 6,1).
>
> Autre exemple, les enfants du groupe « précocité intellectuelle » obtiennent bien des scores sensiblement supérieurs dans les différents indicateurs, mais la différence est moins marquée pour les subtests en lien avec la mémoire de travail, avec au final une moyenne de 107,3 (donc légèrement au-dessus de la moyenne) pour la note composite IVT.

Nous ne pouvons pas ici présenter les conclusions de ces différentes études et renvoyons le lecteur intéressé vers le manuel (manuel 2, p. 63-83).

La validité de la structure du WISC-IV

Comme nous l'avons indiqué à plusieurs reprises, le WISC-IV propose cinq types d'indicateurs de la performance du sujet : un indicateur global (QI T) et quatre indicateurs spécifiques (ICV, IRP, IMT et IVT). Il va s'agir ici de vérifier le bien fondé de ces regroupements de subtests. Par exemple, pour valider le calcul de l'indicateur global QIT, il faut que tous les subtests présentent un certain niveau de liaison entre eux. De même, pour vérifier la validité de structure au niveau des indicateurs spécifiques (les notes composites) on doit vérifier que la liaison de chaque subtest avec son échelle de rattachement (sa note composite) est bien supérieure à sa liaison avec une autre échelle.

1. Il s'agit d'études concernant la précocité intellectuelle, l'épilepsie et la dyslexie.

Analyse du QIT

Pour justifier le calcul du QIT, chaque subtest doit donc présenter une liaison non négligeable avec les autres subtests, liaison qui indique bien que toutes les situations du test évaluent une même dimension, que l'on interprète ici comme l'intelligence globale. Si on consulte le tableau 5.1 du manuel 2 (p. 45) on s'aperçoit que les intercorrélations entre les subtests obligatoires varient de .12 pour la plus faible (entre Mémoire de chiffres et Code) à .67 pour la plus élevée (entre Vocabulaire et Similitudes). Il est tout à fait normal que certains subtests présentent entre eux des valeurs plus élevées de corrélation, explicables par la plus grande proximité de ce qu'ils évaluent, l'essentiel étant d'observer des corrélations significatives entre tous ces subtests. L'existence de telles corrélations valide le calcul d'un indicateur global, le QIT, reposant sur l'ensemble des subtests.

Analyse des notes composites

Il s'agit ici de vérifier la structure du WISC-IV en quatre facteurs. On s'attend ici à observer des saturations maximales de chaque subtest sur son échelle de rattachement. Effectivement, une analyse factorielle exploratoire confirme cette structure. Les données du manuel (p. 45) nous indiquent également que, d'une façon générale, les subtests rattachés à une même note composite corrèlent plus fortement entre eux qu'avec les autres subtests.

On peut signaler ici que l'expérimentation sur l'échantillon français confirme les données américaines, ce qui, rappelons-le, n'était pas le cas lors de l'expérimentation du WISC-III. Le tableau 3.17 présente les saturations observées entre chaque subtest et son échelle de rattachement (ou note composite).

Dans ce tableau, dont les subtests optionnels figurent entre parenthèses, on peut noter que les saturations entre subtests et note composite de rattachement sont assez élevées mais varient selon les indices. Pour l'indice ICV, les valeurs restent élevées (de .67 à .78), de même pour IVT (excepté le .45 observé pour le subtest barrage), par contre les saturations sont plus faibles pour l'indice IRP (de .50 à .56) et pour IMT (.46 à .62).

Cette première étape de validation des quatre indices est confirmée par des analyses supplémentaires, utilisant des méthodes d'analyses factorielles confirmatoires. Rappelons que l'intérêt de ces méthodes confirmatoires est de tester la, ou les, structure(s) factorielle(s) qui a(ont) été définie(s) a priori par les chercheurs.

Il faut se rappeler ici que l'un des objectifs de l'élaboration du WISC-IV, qui a guidé les modifications apportées au WISC-III, était d'obtenir quatre

Tableau 3.17
Saturations factorielles des subtests sur leur échelle de rattachement (analyses factorielles exploratoires, manuel 2, p. 49).

	Les notes composites			
Subtests	**ICV**	**IRP**	**IMT**	**IVT**
Similitudes	.73			
Vocabulaire	.78			
Compréhension	.68			
(Information)	(.70)			
(Raisonnement verbal)	(.67)			
Cubes		.54		
Identification de Concepts		.50		
Matrice		.54		
(Complètement d'images)		(.56)		
Mémoire de chiffres			.56	
Séquence Lettres-Chiffres			.62	
(Arithmétiques)			(.46)	
Code				.69
Symboles				.67
(Barrage)				(.45)

indices fiables du fonctionnement cognitif du sujet. Par exemple, la création du subtest *Matrice* était destinée à renforcer l'indicateur IRP comme mesure de l'intelligence fluide. Les auteurs avaient donc une idée très précise de la structure de l'épreuve, une structure en quatre facteurs (les quatre notes composites), chacun étant obtenu à partir d'une combinaison déterminée de subtests.

C'est ce modèle théorique, défini a priori, qu'il s'agit de tester à l'aide de méthodes confirmatoires. Sans entrer dans les détails de ces analyses[1] nous en retiendrons uniquement ici les principaux résultats obtenus : parmi les différents modèles théoriques testés, c'est bien le modèle postulé (4 facteurs constitués des subtests définis a priori) qui représente le mieux les données observées. La structure du WISC-IV en quatre facteurs est donc confirmée.

Au final l'ensemble des analyses, exploratoires et confirmatoires, valident la pertinence des cinq indicateurs du WISC-IV : QIT, ICV, IRP, IMT et IVT.

1. Nous renvoyons le lecteur intéressé par ces aspects au manuel 2 (pages 51 à 54).

Les bases de l'interprétation du WISC-IV

Les auteurs du manuel rappellent que dans la phase d'interprétation des résultats de l'enfant, le psychologue doit prendre en compte, en plus des performances évaluées au WISC-IV, un ensemble varié d'informations, quantitatives et qualitatives (anamnèse, résultats scolaires, contexte familial et social, comportement pendant la passation...) qui lui permettront de proposer des pistes explicatives au profil des résultats observés.

Ils indiquent également que le WISC-IV se situe dans la tradition des échelles de Wechsler et que, à ce titre, les méthodes et procédures d'interprétation des résultats élaborées pour les autres versions de ces échelles – et nous pensons plus particulièrement ici au WISC-III – sont toujours pertinentes. Le psychologue pourra ainsi consulter avec profit, comme nous l'y avons déjà invité, les parties de cet ouvrage consacrées au WISC-III, ainsi que le dernier ouvrage de Grégoire consacré au WISC-IV (Grégoire, 2006).

Nous avons déjà indiqué que, pour cette version WISC-IV, le psychologue dispose de deux manuels.

Le second manuel (manuel 2) est composé de 123 pages, dont seulement 11 (le chapitre 6) sont exclusivement consacrées à l'interprétation des résultats. Nous aurions aimé que cette partie, essentielle pour le praticien, soit davantage développée. De même, il nous semble que des études de cas mériteraient de figurer dans ce manuel d'autant plus que les modifications sont nombreuses entre le WISC-III et le WISC-IV, ce qui ne va pas faciliter le transfert de compétence que doit opérer le psychologue entre ces deux outils.

Après la cotation de chaque subtest et la transformation des notes brutes en notes standard, le psychologue doit remplir les différentes rubriques du cahier de passation qui lui permettront de procéder à l'analyse des résultats. Pour cela il peut suivre les indications du manuel 1 (pages 49 à 63) qui détaillent les étapes de cette procédure.

➤ Indications préalables

Avant de proposer un cadre général d'interprétation des résultats, les auteurs du manuel passent en revue quelques notions essentielles sur la mesure en général, et sur les indicateurs du WISC-IV en particulier. Nous avons déjà présenté ces notions mais il nous a semblé pertinent de reprendre ces éléments afin de proposer au lecteur une rapide synthèse sur ces notions clefs dont la maîtrise sera nécessaire dans la phase d'interprétation des résultats.

Le psychologue confirmé pourra éventuellement survoler cette partie et aborder directement la partie spécifique concernant le cadre d'interprétation des résultats.

Nous aborderons ici succinctement, les points suivants : la notion de note standard, la métrique des notes composites, la notion de rang percentile, l'erreur type de mesure et l'intervalle de confiance, la description qualitative des notes composites et enfin, les possibilités d'équivalence en niveau d'âge.

La notion de note standard et de rang percentile

La note standard est une note étalonnée, en référence aux performances relevées sur des enfants de même âge.

Au niveau des subtests, les notes standards suivent une loi normale, de moyenne 10 et d'écart-type 3. Ainsi on peut retenir qu'environ 68 % des enfants obtiennent une note standard comprise entre 7 (moyenne – un écart-type) et 13 (moyenne + un écart type). Le tableau 6.1 du manuel (manuel 2, p. 86) permet de connaître le rang percentile de chaque note standard.

> ***Exemple***
>
> À une note standard de 6 correspond un rang percentile de 9, ce qui signifie que seulement 9 % des enfants du même âge ont une note inférieure ou égale à 6.

Au niveau des notes composites, les notes sont exprimées sur une échelle normalisée de moyenne 100 et d'écart-type 15. Le tableau 6.2 du manuel (manuel 2, p. 86) fournit les informations concernant les rangs percentiles de chaque valeur de note composite.

L'erreur-type de mesure

Rappelons que le score observé n'est qu'une estimation de la valeur de la note « vraie » du sujet dans la dimension évaluée et qu'il est plus valide de caractériser le sujet par un intervalle de confiance que par une estimation ponctuelle (voir la présentation de l'erreur de mesure et de la théorie du score vrai dans le chapitre 2 de cet ouvrage). Le psychologue est alors invité à entourer le score observé dans chaque indice (QIT et les quatre notes composites) d'un intervalle de confiance en utilisant les données du manuel 1 (annexe A, p. 237-240).

> ***Exemple***
>
> Pour un risque de 10 %, un QI T observé de 110 devra être entouré de l'intervalle de confiance [104-115]. Autrement dit, on considère que dans 90 % des cas, le score réel (score vrai) du QI T de l'enfant se situe entre 104 et 115.

La description qualitative des notes composites

Comme dans les autres versions des échelles de Wechsler le manuel propose une catégorisation des indices principaux (le QIT et les quatre notes composites) qui reprend d'ailleurs globalement les catégories utilisées dans la WAIS-III : de la catégorie « très faible » (score de 69 et moins) à la catégorie « très supérieur » (score de 130 et plus).

On remarquera ici l'abandon de la dénomination « retard mental », qui figurait dans le manuel du WISC-III pour les scores les moins élevés, au profit ici de la dénomination « très faible » pour les mêmes scores.

Cette grille de catégorisation est reprise en dernière page du cahier de passation, accompagnée des pourcentages de sujets appartenant à chaque catégorie. Le psychologue est invité à situer le niveau de performance de l'enfant de la manière suivante :

> « Par rapport aux enfants du même âge, le fonctionnement intellectuel de cet enfant, mesuré à l'aide d'un test standardisé, se situe actuellement dans la zone [insérer ici la catégorie qualitative appropriée]. » (manuel 2, p. 87)

L'équivalence en niveau d'âge

Le psychologue dispose également de données lui indiquant l'âge moyen auquel une note brute est obtenue dans chaque subtest. Mais en fait, tout en proposant ces références, proches de la notion d'âge mental, le manuel expose les nombreuses limites de leur utilisation et conseille finalement de ne pas les utiliser (manuel 2, p. 88).

Ces différents points ayant été rappelés, nous pouvons maintenant aborder l'interprétation des résultats.

➤ Analyse et interprétation des résultats

Comme dans le cas des autres échelles de Wechsler l'interprétation du profil des résultats est sans doute la partie la plus délicate dans l'utilisation du test, mais aussi la partie la plus intéressante, qui suppose à la fois une bonne maîtrise des concepts théoriques sous-jacents (aspect particulièrement essentiel ici avec cette version WISC-IV pour pouvoir en interpréter les nouveaux indices) mais également de réelles qualités cliniques, de manière à intégrer dans cette phase un ensemble de variables différentes (et de statut différent) : scores aux indices, profil des performances, indications cliniques recueillies pendant la passation.... Cette phase d'interprétation suppose donc, de la part du psychologue, une bonne capacité de synthèse, une bonne maîtrise de l'outil et une expérience clinique dans la passation d'épreuves.

Les étapes de l'analyse des résultats de l'enfant

Le cadre général d'interprétation des résultats du WISC-IV est comparable à celui préconisé dans les autres versions des échelles de Wechsler : il s'agit toujours de débuter l'analyse par l'indicateur global (le QIT), avant d'aborder les indicateurs spécifiques (ici les quatre indices ou notes composites). Dans un premier temps, ces indices seront pris isolément, puis dans un second temps, l'analyse sera centrée sur les différences éventuelles entre les valeurs de ces quatre indices. Enfin, le praticien pourra procéder à l'analyse des mesures les plus spécifiques (les notes aux subtests et, éventuellement, les notes additionnelles).

Comme pour le WISC-III, il s'agira toujours d'articuler deux types de comparaison :

– une comparaison *interindividuelle* (comment l'enfant se situe-t-il par rapport à la réussite moyenne des enfants de son âge ?) ;
– une comparaison *intra-individuelle* (quels sont les domaines sur lesquels il réussit le mieux, comparativement à ses autres résultats ?).

Les deux manuels du WISC-IV fournissent au psychologue un ensemble assez vaste de références diverses concernant l'interprétation des scores de l'enfant : tables d'étalonnage (conversion des scores bruts en notes standard), tables relatives à l'interprétation des différences entre indices, et entre subtests... Ainsi il dispose de près de 70 pages de tableaux divers (pages 204 à 270), composant les annexes A et B du manuel 1, dans lesquels cependant le psychologue non familier du WISC-IV aura sans doute quelques difficultés à s'y retrouver.

Nous avons donc tenté de synthétiser les informations principales concernant cette phase d'interprétation des résultats en reprenant le principe de la décomposition en 10 étapes qui est proposée dans le manuel 2.

Cette partie n'est pas destinée à se substituer à la lecture du manuel mais elle est conçue comme une introduction, une invitation à approfondir les points abordés ici par la lecture des manuels du WISC-IV, ainsi que les publications concernant l'utilisation de cette épreuve (voir en particulier Grégoire, 2006).

Étape 1 : Description du QI Total

Rappelons que le QI T est l'indicateur le plus fidèle, le plus valide et le plus complet de l'ensemble des capacités cognitives de l'enfant. Comme pour le WISC-III, le QIT est un indicateur de l'intelligence globale de l'enfant. Il repose sur les 10 subtests obligatoires. Comme nous l'avons indiqué, ce QI T doit être entouré d'un intervalle de confiance. Le psychologue pourra

nuancer éventuellement cet intervalle à partir de ses observations concernant la passation, comme par exemple le degré d'investissement de l'enfant dans les tâches proposées.

Un premier niveau d'analyse consiste à catégoriser le niveau de performance de l'enfant (de « très faible » à « très supérieur ») à partir des indications du manuel 2 (p. 87). Cette catégorisation permet de situer les résultats d'un enfant par rapport aux enfants de son âge. Pour affiner ce positionnement le psychologue est invité à reporter sur le cahier de passation le rang percentile correspondant.

Enfin, il se pose la question de l'homogénéité de ce score global à partir de l'estimation globale (qui sera plus tard précisée) du profil des résultats : le profil semble-t-il relativement homogène (le QIT étant obtenu à partir de valeurs comparables dans les quatre notes composites) ou le profil semble plutôt hétérogène (cas où on observe une, ou plusieurs, différence(s) sensible(s) entre les notes composites) ?

Cette première estimation du profil des résultats sera ensuite affinée dans les étapes ultérieures à partir de la démarche type, qui constitue une sorte de fil rouge dans l'interprétation des résultats et qui repose sur ces deux questions centrales :

- La différence observée est-elle statistiquement significative ?
- La différence observée est-elle fréquente au sein de la population de référence ?

Après cette analyse globale du QI T, le psychologue va adopter la même procédure (score, intervalle de confiance, classification du score observé, rang percentile, ...) pour analyser les quatre notes composites. Il va s'agir également de préciser ce qui est évalué par chaque indicateur. Ce sont les étapes 2 à 5 que nous allons maintenant aborder.

Étape 2 : Description de l'indice ICV

L'Indice de Compréhension Verbale, qui repose sur trois subtests obligatoires (Similitudes, Vocabulaire et Compréhension), est une mesure du raisonnement à partir de situations dans lequel le langage intervient, mais c'est aussi une estimation de l'intelligence cristallisée qui repose en partie sur les apprentissages réalisés par l'enfant.

Étape 3 : Description de l'indice IRP

L'Indice de Raisonnement Perceptif repose également sur trois subtests obligatoires (Cubes, Identification de concepts et Matrice). Mesure du raisonnement perceptif, il évalue plutôt l'aspect fluide de l'intelligence,

comme en témoigne l'un des nouveaux subtests, Matrice, directement inspiré de tests de type facteur *g*.

Par rapport à l'ancien indicateur QIP, seul un subtest a été conservé : Cubes.

Pour le psychologue familier du WISC-III il s'agit donc ici de bien distinguer ce qui est évalué par ce nouvel indice IRP de ce qui était évalué par le QIP. En particulier, IRP est un indice des capacités de raisonnement à partir de *stimuli* perceptifs et il met moins l'accent sur le raisonnement visuo-spatial (Grégoire, 2005). De plus l'impact de la vitesse de raisonnement est ici réduit (réduction des bonus de temps).

Étape 4 : Description de l'indice IMT

L'Indice de Mémoire de Travail ne repose lui que sur deux subtests obligatoires (Mémoire de chiffres et Séquence Lettres-Chiffres). C'est une mesure moins fidèle que les deux indices précédents.

IMT évalue les capacités de l'enfant à conserver temporairement certaines informations, tout en effectuant un traitement sur celles-ci.

Si le psychologue le souhaite, il peut calculer les notes additionnelles « ordre direct » (plus proche de la notion de mémoire à court terme) et « ordre inverse » (plus proche de la notion de mémoire de travail) de manière à distinguer ces deux facettes de la mémoire.

Étape 5 : Description de l'indice IVT

L'Indice de Vitesse de Traitement ne repose également que sur deux subtests obligatoires (Code et Symboles). Cet indice fournit une évaluation des capacités de l'enfant à traiter rapidement des informations. Nous formulerons deux remarques sur cet indice :

- Si l'on observe les corrélations avec le K-ABC (voir tableau 3.16) on peut remarquer que la corrélation est en réalité plus élevée avec l'échelle de processus simultanés (.41) qu'avec l'échelle des processus séquentiels (.30), ce qui tendrait à montrer que, contrairement à ce que nous indique le manuel (manuel 2, p. 90), l'IVT serait plus proche d'un traitement simultané. Ce point nécessiterait un approfondissement ;
- Le fait qu'un traitement plus rapide (score élevé en IVT) permet de traiter plus d'informations et/ou d'épargner les ressources de la mémoire de travail explique la corrélation souvent observée entre niveau de performance et vitesse de traitement.

Arrivé à la fin de cette cinquième étape, le psychologue a donc caractérisé le sujet sur les indicateurs principaux du WISC-IV. Parmi ces cinq indicateurs

(QIT, ICV, IRP, IMT et IVT), on peut considérer les trois premiers (QIT, ICV et IRP) comme les principaux indicateurs des capacités cognitives globales de l'enfant, IMT et IVT apportant des informations sur des aptitudes plus spécifiques.

Dans les étapes suivantes, le psychologue va s'intéresser aux éventuelles différences entre ces indices.

Rappelons ici deux règles fondamentales dans l'interprétation d'une différence entre deux scores :

- Il convient d'une part de vérifier le caractère significatif, au sens statistique, de cette différence. En effet une différence trop faible, statistiquement non significative, peut résulter de l'erreur de mesure et de fluctuations « normales » (aléatoires), et ne fera pas obligatoirement l'objet d'une analyse ;
- Il convient d'autre part de vérifier la fréquence de cette différence dans la population de référence : une différence, même significative, mais relativement fréquente, ne fera pas non plus l'objet d'interprétations.

Pour pouvoir considérer un profil, ou une partie du profil, comme hétérogène, et donc susceptible d'analyse approfondie, il est donc nécessaire que la, ou les, différence(s) observée(s) soit(ent) à la fois significative(s), au sens statistique du terme, et relativement rare(s). Comme nous l'indiquerons, le manuel fournit les informations nécessaires pour traiter ces deux aspects.

Étape 6 : Comparaison des différences entre les quatre indices principaux (ICV, IRP, IVT et IMT)

Il faut ici analyser les différences entre les indices pris deux à deux, soit 6 comparaisons. Le psychologue dispose pour cela d'un tableau « comparaisons des différences » (tableau situé en haut de la page 2 du cahier de passation) qu'il est invité à compléter. Il note les valeurs des quatre indices, calcule les différences, puis se reporte au tableau B.1 du manuel (manuel 1, p. 256) pour connaître la valeur critique de chaque différence, valeur à partir de laquelle une différence sera statistiquement significative. Le tableau B1 donne ces valeurs critiques en fonction de l'âge de l'enfant et du seuil de confiance retenu (.15 et .05).

Si la différence observée n'est pas supérieure à la valeur critique lue dans le tableau B.1 (au seuil .15), on considère que les résultats de l'enfant sont du même niveau dans les deux indicateurs.

Si la différence observée est égale ou supérieure à la valeur critique (au seuil .15 et éventuellement au seuil .05[1]), il faut alors analyser la fréquence de cette différence sur l'échantillon d'étalonnage afin de pouvoir connaître sa fréquence d'apparition.

Pour cela, le psychologue se réfère aux tableaux B.2 du manuel (manuel 1, p. 257-262), qui lui indiquent le pourcentage d'enfants qui ont obtenu une différence au moins égale aux valeurs indiquées. Attention, il faut utiliser de préférence le tableau correspondant au niveau du QIT de l'enfant (cocher alors la case « par niveau » sur le cahier de passation[2]) et distinguer également le sens de la différence entre les deux indices. Ces informations permettent de remplir la colonne « taux observé » dans le tableau du cahier de passation.

Se pose alors ici le choix d'un seuil critique : à partir de quelle proportion de sujets peut-on considérer une différence comme rare (c'est-à-dire peu fréquente) ? Sattler propose de considérer comme inhabituelle des différences qui ne s'observent que chez moins de 15 % de sujets (manuel 2, p. 91).

Si l'on suit ces indications il faut donc que la valeur de la différence observée soit inférieure à 15 dans ce tableau B.2 pour considérer une différence comme significative mais cette fois au sens *clinique* du terme.

> ***Exemple***
>
> Un enfant de 15 ans qui obtient un QIT de 115, avec un indice ICV de 120 et un indice IRP de 108 aura une différence ICV-IRP de + 12[3].
>
> Le tableau B.1 nous indique que cette différence est statistiquement significative au seuil de .15 (mais elle ne l'est pas au seuil de .05) et le tableau B.2 nous apprend qu'une différence égale ou supérieure s'observe chez 23 % des enfants de même niveau de QIT.
>
> Cette dernière information modère alors l'importance que l'on peut accorder à cette différence : elle n'est pas considérée ici comme inhabituelle car elle s'observe chez près d'un enfant sur quatre et ne donnera donc pas lieu à interprétation.

Bien entendu, il faut aborder ces différents seuils avec souplesse et l'on pourra éventuellement proposer des hypothèses explicatives à propos de différences statistiquement significatives, mais relativement fréquentes dans la population. Dans ce cas, il conviendra d'être plus prudent dans l'interprétation de la signification de cette différence.

1. Le psychologue indiquera si une différence significative à .15 l'est également au seuil de .05.
2. Une autre possibilité de comparaison est possible : prendre en compte l'ensemble de l'échantillon ; cocher alors la case « tout l'échantillon ».
3. Rappelons qu'une valeur minimale de 12 points de différence était nécessaire entre les QIV et QIP du WISC-III pour que l'on considère cette différence comme non négligeable.

Étape 7 : Mise en évidence des forces et des faiblesses

Il va s'agir ici d'analyser les variations intra-individuelles du niveau de performance du sujet à travers ses scores aux différents subtests. Rappelons que l'observation d'un certain niveau d'hétérogénéité des résultats est assez fréquente, un profil réellement plat est, à l'inverse, plutôt rare. Ici encore, comme lors de l'étape précédente, toute différence observée ne mérite pas obligatoirement une analyse précise. C'est en se référant à ce principe général que le psychologue va analyser le profil des résultats du sujet à travers les différents subtests (on parle ici d'analyse du *scatter*) et tenter de repérer le(s) subtest(s) dans le(s)quel(s) il obtient un score plus élevé (ses forces), ou plus bas (ses faiblesses), que sa moyenne personnelle. Il s'agit bien de comparaison intra-individuelle : situer le niveau de performance du sujet dans les différentes tâches du WISC-IV (les différents subtests), non plus par rapport à la moyenne de 10 mais par rapport à la moyenne de ses résultats (sa moyenne personnelle).

Comment calculer cette moyenne personnelle ? Il faut distinguer deux situations :

- Si aucune différence significative entre ICV et IRP n'a été observée (*cf.* l'étape précédente) la moyenne personnelle du sujet est calculée sur les dix subtests obligatoires ;
- Dans le cas inverse, si une différence significative est observée entre ICV et IRP, il faut calculer deux moyennes : l'une à partir des trois subtests obligatoires de l'indice ICV, l'autre à partir des trois subtests IRP. Il faudra alors utiliser comme référence la moyenne de l'indice auquel est rattaché le subtest considéré.

Le psychologue calculera alors les écarts entre le score de chaque subtest et la moyenne personnelle du sujet. Comme pour l'analyse des autres comparaisons, seule une différence statistiquement significative, et relativement rare, pourra être le signe clinique d'un certain niveau de singularité du profil.

Quelques remarques générales concernant cette étape d'interprétation des résultats

1. La mise en évidence éventuelle de force(s) et de faiblesse(s) doit être relativisée par rapport au niveau global de performance : il s'agit bien ici de force(s) et/ou de faiblesse(s) relative(s), qu'il conviendra ensuite de nuancer en fonction du niveau du QI T observé.

2. L'interprétation du profil des résultats du sujet (analyse du scatter), de ses éventuelles forces et faiblesses, doit reposer sur ce qui est évalué dans chaque subtest ou groupe de subtests. Rappelons que pour le WISC-III le psychologue pouvait utiliser une grille d'aide à l'interprétation lui permettant de repérer ce qu'il y avait de commun entre différents subtests (voir Grégoire, 1996). Le psychologue peut éventuellement s'inspirer de cette grille pour l'interprétation des subtests du WISC-IV qui figuraient dans le WISC-III.
3. La méthodologie proposée dans le manuel consistant à choisir entre deux types de comparaisons selon les différences observées entre ICV et IRP (une moyenne générale ou deux moyennes séparées), est en fait à prendre avec souplesse car il est tout à fait possible de procéder aux deux types de comparaison (manuel 2, p. 92).

Étape 8 : Comparaison des différences entre paires de subtests

Cette étape consiste à analyser plus finement les résultats du sujet à partir de ses scores dans deux subtests particuliers. Les tableaux B.3 et B.4 (manuel 1, p. 264-264) indiquent les seuils critiques de différences entre deux subtests, ainsi que les fréquences observées de ces différences dans l'échantillon de sujets de l'étalonnage. L'analyse se fera ici à partir d'hypothèses spécifiques du psychologue et/ou des propositions du manuel. Le cahier de passation comporte ainsi un emplacement pour indiquer trois différences particulièrement intéressantes à analyser :

– mémoire de chiffres/séquence lettres-chiffres ;
– code/symboles ;
– similitudes/identification de concepts.

Étape 9 : Évaluation du profil des notes au sein des subtests

Le psychologue va étudier ici le profil des réussites et des échecs au sein de chaque subtest. Le profil observé doit être conforme à la logique de construction de l'épreuve qui ordonne les items en fonction de leur niveau de difficulté. Ainsi il est rare qu'une réussite apparaisse après plusieurs échecs consécutifs (d'où la validité des règles d'arrêt). Mais certains enfants peuvent présenter un certain degré d'hétérogénéité de leurs résultats à ce niveau.

Par exemple, un enfant brillant peut répondre trop rapidement aux questions faciles (au risque de faire une erreur), tout en réussissant les items suivants, plus difficiles mais, pour lui, sans doute plus intéressants.

Cette analyse des erreurs peut éventuellement permettre de repérer des profils de résultats qui mériteraient une attention particulière.

Étape 10 : Analyse des notes additionnelles

L'une des nouveautés de cette version WISC-IV est la possibilité de bénéficier de scores supplémentaires : les notes additionnelles. Trois subtests sont concernés : Cubes, Mémoire de chiffres et Barrage.

Pour *Cubes*, le psychologue dispose déjà de la note au subtest mais il peut également prendre en compte la note additionnelle « Cubes sans bonifications de temps ». L'observation de la différence éventuelle dans les résultats de l'enfant dans les deux situations doit permettre d'évaluer le poids du facteur vitesse.

Pour *Mémoire de chiffres*, autre subtest obligatoire, le psychologue peut distinguer deux résultats, « mémoire en ordre direct » et « mémoire en ordre indirecte », avec ici aussi la possibilité de comparer ces deux performances.

Enfin, pour le subtest *Barrage*, subtest optionnel, là encore deux mesures : Barrage en ordre aléatoire et Barrage en ordre structuré.

Pour l'analyse de ces différents scores, le psychologue dispose de tableaux (manuel 1) lui permettant de calculer des notes standards, mais il dispose également de références concernant les seuils critiques de signification statistique et des informations sur la fréquence des écarts dans la population d'étalonnage afin de pouvoir effectuer des analyses comparatives entre deux mesures différentes du même subtest. Le manuel donne quelques indications sur la signification de ces différents scores.

Conclusion sur le WISC-IV

Le WISC-IV est la version la plus récente de l'épreuve de Wechsler pour enfant et adolescents, utilisable auprès de sujets âgés de 6 à 16 ans $\frac{1}{2}$. Il s'agit d'une épreuve individuelle, qui repose sur une approche globale de l'intelligence, et qui fournit au psychologue un indicateur concernant le niveau global d'efficience du sujet (le QIT) et quatre indices spécifiques (ICV, IRP, IMT et IVT). Cette nouvelle structure du WISC se rapproche ainsi des modélisations théoriques des aptitudes intellectuelles, principalement celle proposée par Carroll (voir chapitre 1).

Comme nous l'avons indiqué, les modifications sont nombreuses par rapport à l'ancienne version WISC-III, l'une des plus visibles étant sans aucun doute la disparition des deux indices classiques : QIP et QIV. Les utilisateurs du WISC-III risquent d'être, au moins dans un premier temps, assez

désorientés en raison de la difficulté à transférer rapidement leur expérience du WISC-III à l'interprétation des résultats du WISC-IV. D'autant plus, et nous le regrettons, que les manuels ne contiennent aucune illustration d'interprétation de scores ou d'étude de cas (le lecteur trouvera quelques études de cas dans l'ouvrage de Grégoire de 2006).

Certains psychologues regrettent l'importance des modifications apportées au WISC-IV, comme par exemple, la diminution du nombre de subtests qui nécessitent une manipulation concrète du matériel (il ne reste que le subtest Cubes), situations dans lesquelles l'observation de la conduite de l'enfant apportait souvent des éléments pertinents d'information. D'autres encore se questionnent sur ce qui est réellement évalué dans cette version IV (Rozencwajg, 2006).

Par contre, il faut se féliciter de la qualité des études de validation, comme d'ailleurs dans les autres échelles de Wechsler, qui apportent de multiples éléments sur la fiabilité de la mesure, même s'il manque, nous l'avons signalé, des études prédictives par rapport à la réussite scolaire.

Progressivement, le WISC-IV succède dans les pratiques au WISC-III, et sera sans doute l'un des tests d'intelligence le plus utilisé en France.

Cette situation de domination du WISC risque cependant d'être contestée, au moins en France, par l'arrivée de nouvelles épreuves, comme par exemple la Némi-II (Cognet, 2005) ou le K-ABC-II (Kaufman, 2005).

Après la présentation des échelles de Wechsler pour enfant et adolescents, abordons maintenant la version WAIS pour adulte.

4. La WAIS-III

Cette échelle de Wechsler pour adultes version WAIS[1] (âge minimum de 16 ans) reste très proche dans ses fondements théoriques, dans le choix des subtests, comme dans les principes de passation, de cotation et d'interprétation, des échelles WISC-III et WISC-IV pour enfants et adolescents qui viennent d'être présentées.

Tout en reprenant le plan général que nous avons suivi pour la présentation des échelles WISC, ce chapitre sera un peu moins détaillé compte tenu que nombre de propos sur le WISC peuvent être généralisés à la WAIS-III.

1. *Wechsler Adult Intelligence Scale.*

Nous nous appuierons ici essentiellement sur les données du manuel de l'épreuve (Wechsler, 2000) ainsi que sur un ouvrage de Grégoire dans lequel la WAIS-III occupe une place importante (Grégoire, 2004). Signalons également l'ouvrage de Castro (2006) consacré en partie à cette épreuve.

Présentation de l'épreuve

➤ Historique

La première version de cette épreuve américaine, version Wechsler-Bellevue, date de 1939, et est adaptée en France en 1956. La version WAIS est éditée en 1955 et adaptée en France en 1968. Cette version est ensuite révisée en 1981, version WAIS-R, et adaptée en France en 1989. Enfin, la version la plus récente, WAIS-III paraît en 1997, puis est adaptée en France en 2000.

➤ Le matériel

Comme pour le WISC-III, la WAIS-III se présente sous la forme d'une mallette qui contient tout le matériel nécessaire à la passation. Le psychologue dispose d'un manuel bien documenté de 357 pages. Le praticien consigne sur un cahier de passation les réponses du sujet et il dispose d'un document récapitulatif pour reporter l'ensemble des scores.

➤ Les subtests

La WAIS-III présente la même organisation que le WISC-III : un ensemble d'épreuves indépendantes (les subtests) qui sont regroupées en deux sous échelles (une échelle verbale et une échelle de performance).

Au total la WAIS-III comporte 14 subtests, dont 11 proviennent de la précédente version WAIS-R. L'échelle verbale comprend 7 subtests (dont 1 optionnel), l'échelle de performance comprend également 7 subtests (dont 2 sont optionnels). Chaque subtest comporte plusieurs items présentés au sujet selon leur niveau de difficulté.

L'une des grandes modifications par rapport à la version WAIS-R concerne l'introduction d'Indices Factoriels (ce point sera développé plus loin).

Nous présenterons rapidement chacun de ces subtests :

Les subtests de l'échelle verbale

- *Vocabulaire* : consiste à donner la définition de mots (33 items) ;
- *Similitudes* : deux termes sont présentés au sujet qui doit indiquer le type de similitude existant entre ces deux termes (19 items) ;
- *Arithmétique* : petits problèmes arithmétiques à résoudre mentalement (20 items) ;
- *Mémoire des chiffres* : série de chiffres lue au sujet qu'il doit répéter dans le même ordre ou dans l'ordre inverse (8 items en ordre direct, 7 en ordre inverse) ;
- *Information* : questions de connaissances générales (28 items) ;
- *Compréhension* : questions relatives à des problèmes de la vie quotidienne ou de la vie sociale (18 items) ;
- *Séquences lettres-chiffres* : série de chiffres et de lettres, présentées oralement. Le sujet doit les répéter selon un ordre défini : les chiffres, en ordre croissant, puis les lettres, en ordre alphabétique (7 items).

Les subtests de l'échelle de performance

- *Complètement d'images* : le sujet doit indiquer la partie manquante d'un objet, ou d'une situation, représenté par une image (25 items) ;
- *Code* : le sujet doit associer, par écrit, des chiffres à des symboles en respectant des règles d'association ;
- *Cubes* : reproduction de configurations géométriques à l'aide de cubes colorés (14 items).
- *Matrices*[1] : le sujet doit choisir parmi cinq possibilités la matrice qui complète la partie manquante (26 items) ;
- *Arrangements d'images* : série d'images présentées dans le désordre à remettre dans un ordre logique et chronologique (11 items) ;
- *Symboles* : le sujet doit indiquer ici s'il retrouve des symboles cibles au milieu d'autres symboles ;
- *Assemblage d'objets* : sorte de puzzle à reconstruire (5 items).

1. Tâche inspirée des Progressives Matrices de Raven.

Standardisation

➤ La passation

La passation est individuelle et nécessite un temps d'environ 1 heure à 1 heure 30 minutes (durée variable en fonction du niveau de réussite du sujet). La passation des subtests est effectuée selon un ordre déterminé, avec alternance entre les subtests verbaux et les subtests de performance afin de préserver chez le sujet un certain niveau de motivation. Certains subtests sont en temps limité (utilisation alors d'un chronomètre), d'autres en temps libre. Le manuel indique très précisément les règles de passation pour chaque subtest.

Dans certains subtests, la passation ne débute pas au premier item mais à un item donné (par exemple, le quatrième) avec administration des premiers items si le sujet échoue aux deux premiers items proposés. Ces règles de départ spécifiques à chaque subtest permettent d'éviter de faire passer à tous les sujets les items de faible niveau de difficulté tout en conservant ces premiers items pour différencier les sujets les plus faibles. Selon la même logique, des règles d'arrêt sont indiquées pour chaque subtest. Ces règles de départ et d'arrêt permettent ainsi de ne faire passer au sujet que les items les plus en relation avec son niveau de compétence et de réduire le temps de passation.

On peut également noter que le praticien peut décider de ne faire passer qu'une partie des subtests s'il ne souhaite pas recueillir tous les indicateurs de cette échelle (voir plus loin le tableau 3.18).

➤ La cotation et les indicateurs de la performance du sujet

On va retrouver ici la même logique de cotation que celle suivie dans les versions WISC.

La première étape consiste à effectuer la cotation des items de chaque subtest (en 0/1 point ou en 0/1/2 points selon les cas), avec d'éventuels bonus en fonction du temps de réalisation, puis d'en faire la somme. Chaque total est une note brute qui devra ensuite être transformée en note standard (note étalonnée) en fonction de l'âge du sujet. En effet, comme dans toutes les échelles de Wechsler, la population de référence est constituée des sujets de même âge. Comme pour le WISC, les notes standards sont des notes étalonnées de 1 à 19 (moyenne de 10 et écart-type de 3).

À partir des notes standards il est possible alors de calculer sept indicateurs de la performance du sujet :

– les trois Q.I. classiques (QIV, QIP et QIT),
– les quatre Indices Factoriels : Indice Compréhension Verbale (I.C.V), Indice Organisation Perceptive (I.O.P), Indice Mémoire de Travail (I.M.T) et Indice Vitesse de Traitement (I.V.T).

On peut remarquer ici la proximité de la structure de la WAIS-III avec la structure du WISC-IV, même si ce dernier ne comporte plus les QIV et QIP.

Le tableau 3.18 nous indique les rattachements des subtests aux différents indicateurs.

Tableau 3.18
Rattachement des subtests de la WAIS-III aux indicateurs globaux (les parenthèses signalent les subtests optionnels).

	Les 3 Q.I.			Les 4 indices factoriels			
Subtests	**QIV**	**QIP**	**QIT**	**I.C.V**	**I.O.P**	**I.M.T**	**I.VT**
Vocabulaire	X		X	X			
Similitudes	X		X	X			
Arithmétique	X		X			X	
Mémoire des chiffres	X		X			X	
Information	X		X	X			
Compréhension	X		X				
Séquence Lettres-chiffres	(X)		(X)			X	
Complètement d'images		X	X		X		
Code		X	X				X
Cubes		X	X		X		
Matrices		X	X		X		
Arrangement d'images		X	X				
Symboles		(X)	(X)				X
Assemblage d'objet		(X)	(X)				

On peut observer dans ce tableau que les sept indicateurs ne reposent pas tous sur le même nombre de subtests :

– les deux QI (QIV et QIP) reposant respectivement sur 6 et 5 subtests ;
– les Indices Factoriels reposant sur 3 subtests (sauf I.V.T qui ne repose que sur 2 subtests) ;

– le QIT, seul indicateur global de la performance du sujet, qui repose sur l'ensemble des subtests.

Rappelons que tous ces indicateurs sont exprimés dans la même métrique (moyenne de 100 et écart-type de 15) et qu'il convient d'entourer chaque valeur observée par un intervalle de confiance.

Cette version WAIS-III propose une certaine souplesse dans la passation, le praticien ne pouvant faire passer qu'une partie des subtests en fonction des indicateurs qu'il souhaite calculer :

- S'il souhaite obtenir un maximum d'information, et calculer alors Q.I et Indices Factoriels, la passation de l'ensemble des subtests (hormis les optionnels) est nécessaire ;
- S'il ne souhaite obtenir que certains indicateurs (par exemple, uniquement les Indices Factoriels), la passation ne concernera alors qu'une partie des subtests.

Bien entendu, la durée de passation dépendra de ces choix.

Le praticien reportera l'ensemble des résultats sur un document séparé intitulé « récapitulatif/profil ».

➤ Les étalonnages disponibles

Comme pour le WISC, la composition de l'échantillon de sujets constituant l'étalonnage est soigneusement décrite dans le manuel : un échantillon de 1 104 sujets, âgés de 16 à 89 ans, représentatif de la population française (répartition selon la CSP, l'âge, le sexe...). L'étalonnage a été réalisé en 1998/1999.

Douze groupes d'âges ont été constitués afin d'élaborer des étalonnages par classe d'âge.

Les étalonnages concernent les notes standards, les Q.I. et les Indices Factoriels. Le tableau 3.19 permet de résumer les valeurs caractéristiques des différents indicateurs étalonnés de la WAIS-III.

Tableau 3.19

Valeurs caractéristiques des indicateurs de la WAIS-III.

Indicateurs	Valeur Mini	Valeur maxi	Moyenne	Écart type
Notes standards	1	19	10	3
QIV, QIP et QIT	45	155	100	15
IOP, ICV, IMT, IVT	50	150	100	15

Sachant que ces indicateurs suivent une répartition normale on peut considérer que les étalonnages disponibles sont bien adaptés à la population visée.

De plus, pour chaque indicateur, le praticien dispose également de deux types d'information : intervalle de confiance et rang percentile.

Les qualités psychométriques

Nous aborderons ici l'analyse des informations du manuel concernant la sensibilité, la fidélité et la validité de la WAIS-III.

➤ La sensibilité

La sensibilité d'un test représente sa capacité à différencier les sujets. Il s'agit ici de s'assurer que les différents scores permettent bien de discriminer les sujets. Nous pouvons vérifier sur le tableau 3.19 que cette différenciation est assurée par l'étendue des différents indicateurs étalonnés (notes standards, Q.I. et Indices Factoriels) et par la répartition gaussienne de ces scores.

> ***Exemple***
>
> Au niveau des notes standards, elles peuvent varier de 1 à 19 (moyenne de 10 et écart-type de 3) et couvrent ainsi trois écarts type de part et d'autre de la moyenne, ce qui assure un bon niveau de différenciation des sujets. On observe cette même qualité de différenciation au niveau des Q.I. (variation possible de 45 à 155) comme au niveau des Indices Factoriels, même si ces derniers présentent une étendue légèrement plus réduite (de 50 à 150).

L'ensemble de ces données assure à la WAIS-III un bon niveau de différenciation des sujets.

➤ La fidélité

Rappelons que la fidélité concerne la précision et la stabilité de la mesure. Différents types de fidélité sont analysés dans le manuel de la WAIS-III (d'après Wechsler, 2000) :

– *La fidélité*, ou consistance, interne, est évaluée par la méthode pair-impair. Les coefficients de fidélité varient de .68 à .90 selon les subtests, de .92 à .97 selon les Q.I. et de .86 à .95 selon les Indices Factoriels. La valeur de ces indicateurs de fidélité est très satisfaisante ;

– *La fidélité temporelle* a été vérifiée par la méthode test-retest, sur un échantillon de 103 sujets, avec un intervalle de 2 à 13 semaines entre les passations. Les coefficients sont ici encore satisfaisants avec, par exemple, des valeurs comprises entre .86 et .94 pour les Q.I. ;
– *La fidélité intercotateurs*, évaluée sur des subtests verbaux dans lesquelles la cotation est plus complexe (0, 1 ou 2 points par item), est presque parfaite avec des valeurs supérieures à .92 ;
– *L'erreur-type de mesure* est calculée pour chaque type de score : notes aux subtests, notes de Q.I. et Indices Factoriels.

À partir de ces valeurs, il est possible de calculer les intervalles de confiance qui entourent les scores calculés. Le manuel facilite ici la tâche du psychologue en donnant directement les valeurs de ces intervalles pour chaque valeur observée aux seuils .10 et .05.

Si l'on prend, par exemple, une même valeur observée de 103[1], le tableau 3.20 nous indique l'intervalle de confiance, au seuil .10, pour chaque indicateur.

Tableau 3.20
Exemples d'intervalles de confiance pour une valeur observée de 103 (d'après le manuel WAIS-III).

	Intervalle de confiance (seuil de .10)						
	QIV	QIP	QIT	ICV	IOP	IMT[a]	IVT[b]
Valeur observée de 103	98-107	96-109	99-107	97-108	96-109	97-109	95-111

a. La valeur 103 n'existant pas pour cet indice nous avons pris les valeurs intermédiaires situées entre celles correspondant à un score observé de 102 et celles correspondant à un score observé de 104.
b. Même remarque.

On peut remarquer sur ce tableau que l'intervalle de confiance le plus réduit, donc l'erreur de mesure la plus faible, concerne le QIT, avec un intervalle de 9 points [99-107]. En effet, le QIT est la mesure qui porte sur le plus grand nombre de subtests, ce qui explique sa plus faible valeur d'erreur de mesure. L'intervalle sur QIP est un peu plus élevé que celui portant sur QIV, et les intervalles pour les Indices Factoriels sont globalement du même ordre que ceux relatifs aux Q.I. C'est l'Indice I.V.T qui présente

1. Bien entendu il s'agit d'un cas fictif : un même sujet ne présente généralement pas le même score sur les différents indicateurs.

l'intervalle le plus important. Nous avons observé dans le tableau 3.18 que cet Indice I.V.T n'est constitué que de deux subtests, ce qui peut expliquer cette faiblesse.

➤ La validité

Cette qualité fondamentale d'un test peut se résumer en la question suivante : le test mesure-t-il ce qu'il est censé mesuré ?

Pour répondre à cette question, deux aspects principaux seront ici analysés :

- La WAIS-III est-elle une mesure de l'intelligence ?
- Quelle est la validité des différents indicateurs de ce test (Q.I. et Indices Factoriels) ?

➤ La validité de la WAIS-III comme mesure de l'intelligence

Il s'agit ici d'analyser les corrélations observées entre les scores obtenus, par les mêmes sujets, à la WAIS-III et à d'autres tests d'intelligence.

Nous partons ici avec un *a priori* largement positif : comme nous l'avons déjà indiqué, la validité des échelles de Weschler comme mesures de l'intelligence n'est plus à démontrer. Le manuel présente un grand nombre d'études de validation, mais nous pouvons regretter que la plupart portent sur des populations américaines et/ou sur l'ancienne version de cette échelle (WAIS-R). Nous ne présenterons ici que les résultats des études qui concernent l'échantillon français et la version WAIS-III.

Liaison entre WAIS-R et WAIS-III

Une première étape de l'analyse de la validité de la WAIS-III consiste à vérifier que cette épreuve mesure bien les mêmes dimensions que l'ancienne version WAIS-R. Une étude est présentée dans le manuel qui porte sur un échantillon assez faible de sujets (55 sujets). L'analyse des résultats montre que, au niveau des Q.I., les coefficients de corrélation varient de .86 à .93, valeurs qui confirment que la WAIS-III évalue bien les mêmes dimensions que la WAIS-R (intelligence *globale*, intelligence *verbale* et intelligence *non verbale*). Pour les Indices Factoriels, comme ils n'existaient pas dans la version WAIS-R, cette analyse n'est pas possible.

Liaison avec le WISC-III

Nous avons déjà présenté, dans la partie concernant le WISC-III, les valeurs des coefficients de corrélation entre WISC-III et WAIS-III (voir tableau 3.5). Rappelons que les valeurs se situaient autour de .90 pour les Q.I. et entre .76 et .88 pour les deux Indices Factoriels communs (I.C.V et I.O.P). Ces valeurs élevées témoignent de la proximité de ces deux épreuves.

Liaisons avec le WISC-IV

Bien entendu, compte tenu de l'antériorité de la WAIS-III par rapport au WISC-IV, le manuel de la WAIS-III ne comporte aucune donnée à ce niveau mais nous disposons de résultats dans le manuel du WISC-IV. Nous avons déjà présenté ces résultats (voir en particulier le tableau 3.15) qui confirment la proximité de ce qui est évalué par ces deux épreuves : .83 au niveau du QIT, et des valeurs comprises entre .64 et .81 pour les indices factoriels/notes composites.

On peut remarquer qu'il aurait été intéressant de disposer d'études, sur un échantillon français, relatives aux liaisons entre la WAIS-III et un test de type facteur *g*.

➤ La validité des indicateurs de la WAIS-III (Q.I. et Indices Factoriels)

Lorsqu'un test, comme ici, propose de calculer différents indicateurs des performances du sujet, il convient de vérifier les bases sur lesquelles reposent ces indicateurs (notion de validité structurale). Dans le cas de la WAIS-III, la logique de validation des indices sera comparable à celle évoquée dans l'analyse de la validité structurale du WISC :

- La distinction proposée entre les deux échelles, donc le calcul séparé de deux scores (QIV et QIP), doit être justifiée par l'observation de fortes liaisons entre subtests d'une même échelle ;
- Le calcul d'un indice total, le QIT, doit reposer sur le fait que tous les subtests évaluent bien une même dimension commune ;
- Le calcul des Indices Factoriels, doit lui aussi être validé par des méthodes statistiques appropriées. D'autant plus que ces Indicateurs représentent une nouveauté par rapport à la WAIS-R.

Concernant le premier aspect, le manuel donne les résultats d'analyses factorielles descriptives qui valident la distinction classique entre les deux échelles, donc le calcul séparé des deux Q.I : QIV et QIP. En effet, les corrélations entre les subtests de l'échelle Verbale sont bien plus élevées

que les corrélations entre ces subtests et ceux de l'échelle de Performance (Wechsler, 2000, p. 262). Par contre, le manuel indique aussi que cet effet est moins marqué pour les subtests de l'échelle de Performance, en particulier pour *cubes* et *matrices* qui présentent des corrélations assez élevées avec certains subtests de l'échelle Verbale.

Ces analyses montrent également que tous les subtests évaluent bien une même dimension, que l'on interprète ici comme étant *un facteur général d'intelligence*, ce qui permet de valider le calcul du QIT.

Enfin pour l'analyse des Indices Factoriels, leur nouveauté mérite que l'on détaille un peu plus les éléments de leur validation.

Tout d'abord, il faut indiquer l'origine de ces Indices. De manière comparable aux évolutions du WISC, les auteurs de la WAIS-III ont souhaité intégrer dans cette nouvelle épreuve les résultats des recherches les plus récentes dans le domaine de l'intelligence et du fonctionnement cognitif afin d'évaluer plus précisément la mémoire de travail et la vitesse de traitement. Cet objectif explique l'apparition de nouveaux subtests dans la version WAIS-III.

Plus précisément, suite aux résultats de différentes études portant sur la WAIS-R et sur le WISC-III, les auteurs souhaitent obtenir, pour la WAIS-III, une structure comportant quatre facteurs. Ces quatre facteurs doivent correspondre à des mesures spécifiques définies comme :

- la Compréhension Verbale (I.C.V),
- l'Organisation Perceptive (I.O.P),
- la Mémoire de Travail (I.M.T),
- la Vitesse de Traitement (I.V.T).

Pour valider cette structure hypothétique ils créent de nouveaux items, de nouveaux subtests, puis utilisent une méthode d'analyse factorielle confirmatoire, méthode qui permet de tester l'adéquation d'un modèle théorique (modèle composé ici des quatre facteurs) à partir des données observées. Effectivement, l'analyse des résultats confirme cette structure hypothétique en quatre facteurs et valide ainsi le calcul des quatre Indices Factoriels représentant ces quatre facteurs (le lecteur intéressé pourra consulter les pages 270 à 274 du manuel qui traitent spécifiquement de ces analyses).

Nous pouvons reprendre ici la remarque concernant le calcul de ces Indices Factoriels, que nous avons déjà formulé lors de la présentation du WISC-III. En effet, comme c'était aussi le cas dans le WISC-III, les valeurs des liaisons (des saturations) entre subtests et Indice Factoriel varient selon les subtests et un calcul pondéré, tenant compte de ces variations, serait

plus proche des données, donc plus valide, que la simple addition des notes standard des subtests concernés.

Par exemple, le tableau 6.7 du manuel (Wechsler, 2000, p. 266) indique que pour l'Indice Factoriel I.M.T, la saturation est de .76 avec le subtest *Séquence lettres-chiffres* mais seulement de .42 avec le subtest *Arithmétique*. Pourtant, dans le calcul de cet indice, on accorde le même poids à ces deux subtests. Une autre possibilité aurait pu être envisagée qui consisterait à pondérer chaque subtest en fonction de la valeur de sa saturation.

Enfin, toujours à propos de ces Indices Factoriels, Grégoire présente les résultats d'analyses complémentaires qui confirment la validité de cette structure factorielle. Cette décomposition des résultats de la WAIS-III en quatre Indices Factoriels lui paraît même préférable à l'utilisation des deux indicateurs classiques QIV et QIP, car ces indices représentent des mesures plus robustes et plus homogènes que les deux Q.I. classiques (2004, p. 207).

Les bases de l'interprétation

Avec la WAIS-III, le praticien obtient plusieurs indicateurs quantitatifs :

- les notes standards,
- les trois Q.I.,
- les quatre Indices Factoriels.

Il dispose éventuellement d'indices plus qualitatifs relevés lors de la passation de l'épreuve :

- implication du sujet dans les tâches proposées,
- stratégies de résolution,
- comportement face à une difficulté,
- niveau de fatigabilité...

Enfin, à travers les entretiens il peut éventuellement recueillir des informations diverses sur le sujet (diplômes, expériences professionnelles...).

L'étape suivante va donc consister à tenter de synthétiser toutes ces informations afin de mieux comprendre le fonctionnement cognitif de l'individu singulier qui a passé la WAIS-III.

Nous traiterons dans cette partie principalement de l'interprétation des données quantitatives relatives aux différents scores observés à la WAIS-III. Le principe général d'analyse et d'interprétation des résultats à la WAIS-III suit la même logique que celle qui régit l'analyse des résultats au WISC : partir

du général pour se diriger vers le particulier. Il va donc s'agir d'analyser l'indicateur le plus général (le QIT) puis les indicateurs spécifiques (QIV, QIP et les Indices Factoriels) et enfin d'analyser les résultats aux différents subtests.

Avant de présenter les différentes étapes de l'analyse, il faut rappeler que tous les indicateurs étalonnés de la WAIS-III se réfèrent aux performances observées chez des sujets de même âge. Ce point est très important à rappeler, surtout dans le cas où le sujet est relativement âgé. En effet, avec le phénomène de déclin de certaines aptitudes avec l'âge, un sujet de 70 ans ayant un Q.I. de 100 aura en fait un niveau de performance moins élevé qu'un sujet de 30 ans qui a pourtant le même Q.I. de 100. Ces deux sujets se situent de la même manière (ici très précisément au centre de la distribution, au niveau de la moyenne) mais dans des populations de référence différentes. Prenons par exemple deux subtests particulièrement sensibles à ce phénomène de déclin, le subtest *mémoire des chiffres* et le subtest *matrices.* Le tableau 3.21 donne les notes étalonnées (notes standards) pour un même niveau de réussite (score brut) en fonction du groupe d'âge.

Tableau 3.21
Comparaison des différentes notes standard attribuées à un même score brut selon la classe d'âge (d'après Wechsler, 2000, p. 302 à 307).

	Notes standards selon le groupe d'âge			
Score brut observé	20-34 ans	55-64 ans	70-74 ans	80-89 ans
Matrices : 21 points	10	12	16	17
Mém. chiffres : 17 points	10	12	13	14

Les données illustrent bien le phénomène que nous voulions décrire : les notes étalonnées (notes standards) dépendent bien du niveau de réussite observé dans chaque classe d'âge.

Exemple

Pour un même niveau de réussite au subtest *matrice* (un score brut de 21 points), le sujet sera situé juste dans la moyenne s'il est âgé de 25 ans (avec une note standard de 10) mais plus il sera âgé, plus sa note standard sera élevée, avec ici par exemple une note standard de 17 s'il est âgé de 80 ans. On peut également observer un processus équivalent pour l'autre subtest (*mémoire des chiffres*).

Cet effet de variation des niveaux de performances selon les classes d'âge peut intervenir pour tous les subtests et pour tous les indicateurs qui sont calculés à partir de ces notes standards (Q.I. et Indices Factoriels). Il convient donc, avant toute interprétation des résultats, de prendre en compte les notes étalonnées du sujet (qui situent ses performances par

rapport aux sujets de même âge) mais également les valeurs de référence proposées (le groupe d'âge 20-34 ans), surtout si le sujet est éloigné de cette classe d'âge. C'est pour ces raisons qu'il est conseillé de faire figurer sur le document « récapitulatif /profil » les valeurs des notes standard pour le groupe 20-34 ans qui correspond à un groupe de référence éventuellement différent du groupe d'âge du sujet (voir la colonne réservée à cet effet en troisième page de ce document). Enfin rappelons qu'il est fortement conseillé d'entourer chaque score calculé d'un intervalle de confiance.

Abordons maintenant les différentes étapes d'analyse des résultats.

➤ Étape 1 : Analyse du Q.I. Total (QIT)

Cet indicateur QIT est, comme dans toutes les versions des échelles de Wechsler, l'indicateur le plus complet de la WAIS-III, car il prend en compte un grand nombre de subtests, donc un ensemble varié de situations.

Cet indicateur de l'intelligence globale permet de situer le sujet dans une population de référence à l'aide du rang percentile.

> ***Exemple***
>
> Un QI de 109 correspond au rang percentile 73 ce qui signifie que 73 % des sujets obtiennent un score inférieur ou égal à 109 et donc que seulement 27 % des sujets obtiennent un score supérieur.

Tableau 3.22
Classification des Q.I. et des Indices Factoriels au test WAIS-III (d'après Wechsler, p. 280).

Q.I ou Indice Factoriel	% théorique de sujets	Classification (catégorie)
130 et plus	2,2 %	Très supérieur
120-129	6,7 %	Supérieur
110-119	16,1 %	Moyen supérieur
90-109	50 %	Moyen
80-89	16,1 %	Moyen inférieur
70-79	6,7 %	Limite
69 et moins	2,2 %	Très faible

Ce positionnement précis de la performance du sujet peut également être interprété de manière plus qualitative à partir de la classification proposée dans le manuel.

Cette classification, qui concerne plus largement tous les Q.I. et tous les Indices Factoriels, peut être utilisée par le psychologue pour situer de manière plus qualitative le niveau de performance du sujet.

Les valeurs des différents seuils qui déterminent les classes ont été définies en fonction de la répartition théorique des sujets (par exemple, la valeur de 130 correspond à une performance située deux écarts types au-dessus de la moyenne). Sur la justification de ces seuils, nous renvoyons le lecteur aux réflexions que nous avons proposées dans la présentation des WISC-III et IV.

➤ Étape 2 : Analyse du QIV, du QIP et des Indices Factoriels

Il faut, dans un premier temps, rappeler la signification de ces différents indicateurs, puis, dans un second temps, donner les éléments essentiels pour leur analyse.

Que représentent ces indicateurs ?

Pour les deux Q.I. nous pouvons considérer qu'ils correspondent globalement aux indicateurs QIV et QIP du WISC-III, que nous avons déjà présenté (voir présentation du WISC-III). Ainsi le QIV est une mesure du raisonnement verbal mais aussi une évaluation des connaissances acquises. Le QIP étant quant à lui plutôt une évaluation des capacités de raisonnement dans des situations nouvelles, dans lesquelles le langage n'intervient pas, ou peu. Le QIV est considéré comme proche de la notion d'intelligence cristallisée, le QIP étant associé à la notion d'intelligence fluide. Ces deux Q.I. sont également très proches des indicateurs QIV et QIP de l'ancienne version WAIS-R. On peut signaler que, suite à l'introduction du subtests *matrice*, et au fait que le subtest *assemblage d'objet* devient optionnel, le QIP de la WAIS-III devient moins sensible à la vitesse de traitement et plus proche du raisonnement abstrait non verbal.

Pour les Indices Factoriels, nous avions questionné dans la version WISC-III, l'intérêt de ces Indices qui étaient très (trop ?) proches des indicateurs classiques de Q.I. et n'apportaient pas alors d'informations suffisamment spécifiques et/ou suffisamment fiables. Par contre, dans cette version WAIS-III, les Indicateurs Factoriels présentent des différences plus importantes avec les deux Q.I., et constituent des mesures plus « pures » ou plus « fines » (pour reprendre les termes du manuel) du fonctionnement intellectuel du sujet.

Ainsi l'Indice ICV est, du fait de l'absence des subtests *compréhension, mémoire des chiffres* et *arithmétique,* « une mesure plus pure de la compréhension verbale » (Wechsler, 2000, p. 284),

De même, IOP peut être défini comme une mesure plus pure de l'intelligence fluide.

Enfin, les Indices Factoriels IMT de IVT, sont des mesures assez spécifiques, qui apportent des éléments complémentaires sur deux aspects de fonctionnement intellectuel : la mémoire de travail et la vitesse de traitement des informations.

Après avoir situé le niveau de performance du sujet dans chaque indicateur, le psychologue analysera, comme pour le WISC, le profil des résultats du sujet autour de la question suivante : le profil est-il homogène (cas d'une faible différence entre les indicateurs) ou hétérogène (différence importante entre les indicateurs) ?

Attention !

Comme pour le WISC, toute différence observée n'est pas obligatoirement significative.

En effet, pour que cette différence ait un sens au niveau du fonctionnement cognitif il est nécessaire qu'elle soit à la fois assez importante (statistiquement significative) et relativement rare. Les valeurs de référence fournies dans le manuel permettent de guider l'interprétation des différences éventuellement observées.

Par exemple, pour la WAIS-III, la différence moyenne entre QIV et QIP est proche de 10 points (9,7 points) et près de 20 % des sujets présentent une différence égale ou supérieure à 16 points (p. 320).

Ces informations relativisent ainsi grandement la singularité des profils qui présenteraient une différence entre QIV et QIP inférieure ou égale à ces valeurs.

Le manuel propose deux exemples d'interprétation des différences observées, l'un concerne une différence entre les deux Q.I., l'autre une différence entre deux Indices Factoriels (voir p. 289 et 290).

En conclusion, on ne peut que conseiller au praticien de se référer aux informations du manuel (valeurs significatives des différences, répartition de ces différences dans l'échantillon de référence, exemples d'interprétation de profils...) avant d'effectuer toute interprétation des différences observées.

Enfin, on peut rappeler que Grégoire est plus favorable à l'analyse des Indices Factoriels qu'à l'analyse traditionnelle des deux Q.I. (QIV et QIP), les Indices ayant une fiabilité plus importante :

> « Avec la WAIS-III, le calcul des Indices apparaît comme une option préférable au calcul des traditionnels QI Verbal et QI Performance. Les Indices apparaissent en effet comme des mesures plus robustes et plus homogènes que les QI. » (Grégoire, 2004, p. 207).

➤ Étape 3 : Analyses des subtests

La dernière étape de l'analyse concerne l'analyse des résultats aux différents subtests. Rappelons que l'on parle alors d'analyse de *scatter*, c'est-à-dire d'analyse de la dispersion des notes standard. Il s'agit ici d'analyser l'homogénéité du profil des notes standard, de repérer les subtests particulièrement réussis et ceux particulièrement échoués...

Il faudra, comme pour le WISC, articuler deux types de comparaison :

- Une comparaison *interindividuelle*, avec comme référence le niveau moyen de réussite dans la population de référence : il s'agira alors de situer le score observé à chaque subtest par rapport à la valeur moyenne de 10 ;
- Une comparaison *intra-individuelle*, avec comme référence ici la moyenne propre du sujet : il s'agira alors de situer chaque score par rapport à la moyenne individuelle du sujet (moyenne calculée à partir de ses différentes notes standards).

Pour effectuer la comparaison interindividuelle, le praticien peut utiliser la classification des notes standard que nous avons proposée dans la présentation du WISC-III pour repérer les points forts et les faiblesses du sujet (voir tableau 3.10).

Pour l'analyse intra-individuelle, le manuel propose les valeurs seuils des différences significatives notes standard. À partir de ces informations le praticien peut repérer les notes déviantes, qui s'écartent significativement de la moyenne personnelle du sujet et qui peuvent singulariser son profil et son fonctionnement cognitif

Dans la feuille de synthèse des résultats du sujet il est d'ailleurs demandé d'indiquer la valeur de la différence observée entre chaque subtest et la moyenne individuelle, ainsi que le niveau de significativité statistique de ces différences.

Une analyse plus approfondie du protocole, consistant en la prise en compte des covariations des notes à différents subtests, est également possible mais le praticien ne dispose pas ici, contrairement au WISC-III, d'une grille d'aide spécifique (Grégoire, 1996). Mais, compte tenu des nombreux subtests communs entre WISC-III et WAIS-III, il nous semble possible

d'appliquer, au moins en partie, cette grille d'aide à l'interprétation des scores à la WAIS-III.

Conclusion sur la WAIS-III

La WAIS-III, échelle de Wechsler pour adulte, qui repose sur une conception globale de l'intelligence, permet donc au final de disposer de sept indicateurs de l'efficience du sujet :

– les trois Q.I. classiques : QIT, QIV et QIP ;
– les quatre Indices Factoriels : ICV, IOP, IMT et IVT.

Les données disponibles confirment la fiabilité de ce test et de ses différents indicateurs. Pour cette version WAIS-III, et contrairement aux réserves que nous avions formulées pour le WISC-III, les Indices Factoriels apportent bien ici des informations spécifiques sur le fonctionnement cognitif du sujet. D'ailleurs, pour certains auteurs (Grégoire, 2004), en raison d'une fiabilité plus élevée, l'utilisation de ces Indices est préférable à l'utilisation classique des QIV et QIP. Il est probable d'ailleurs que dans la prochaine version (WAIS-IV) le praticien ne dispose plus de ces indicateurs QIV et QIP, mais uniquement des indices factoriels, accompagnés du QIT, comme cela est déjà le cas dans la version WISC-IV.

Il serait intéressant de connaître la proportion de praticiens qui conserve une utilisation traditionnelle de la WAIS-III (analyse préférentielle du QIV et du QIP) et celle qui privilégie l'utilisation et l'interprétation de ces Indices Factoriels. Il est fort probable que ce changement de pratique demande un peu de temps...

Enfin, signalons qu'il n'existe pas de version abrégée de cette épreuve[1] (comme c'était le cas pour le WISC-III) mais qu'une certaine latitude est laissée au praticien lors de la passation, lui permettant, en fonction des indicateurs qu'il souhaite obtenir (les Q.I. et/ou les Indices Factoriels), de ne faire éventuellement passer qu'une partie des subtests de l'échelle.

1. De telles versions font l'objet de recherche (voir par exemple Rémy, 2008).

CHAPITRE
4

Les tests de facteur g (et d'intelligence fluide)

Sommaire

DANS le chapitre 1, nous avons décrit l'apport de Spearman et ses propositions, formulées il y a près d'un siècle, concernant le facteur *g* et la place prépondérante qu'il occupe dans le modèle factoriel de l'intelligence. On en trouve toujours la marque dans les modèles actuels de l'intelligence : le facteur *g* correspond au troisième niveau (niveau supérieur) dans le modèle hiérarchique de l'intelligence de Carroll et est très proche de la notion d'intelligence fluide (Gf), à laquelle se réfèrent de nombreux auteurs.

L'importance et la pertinence du facteur *g* sont largement reconnues dans l'explication des capacités cognitives individuelles, en particulier lorsque l'on cherche à prédire la capacité d'une personne à résoudre des problèmes logiques dans des situations et des domaines variés. Mesurer cette dimension est donc particulièrement utile. C'est l'objectif des tests dits « de facteur *g* » que nous allons maintenant présenter.

Il existe un assez grand nombre d'épreuves de facteur *g* qui, malgré leurs spécificités, présentent de nombreux points communs :

- Ce sont généralement des épreuves prévues pour des passations collectives (avec cependant presque toujours une possibilité de passation individuelle) ;
- Le temps de passation est souvent court, de 15 à 30 minutes ;
- Le niveau de performance du sujet est, le plus souvent, un indicateur unique ;
- Les tâches proposées sont généralement de type « lois de séries » dans lesquelles le sujet doit analyser la situation afin de découvrir les relations (la ou les lois de transformation) qui relient les différents éléments de cette situation, puis appliquer cette loi afin de trouver (ou de sélectionner) la bonne réponse (voir dans le chapitre I les notions d'éduction de relations et de corrélats, proposées par Spearman) ;
- Enfin, dernier point, la part du facteur verbal n'intervient pas ou peu dans ces épreuves.

Nous analyserons ici de manière détaillée les épreuves de facteur *g* les plus connues et/ou les plus utilisées en France :

– les matrices de Raven,
– le test NNAT,

– les épreuves de dominos (D70 et D2000),
– les tests de raisonnement R85 et R2000.

En fin de chapitre nous présenterons aussi, mais de manière plus synthétique, d'autres épreuves du même type également disponibles en France (épreuve de Cattell, BLS4, B53 et RCC).

1. Les tests de Raven

Les épreuves *Progressive Matrices* de Raven sont des exemples prototypiques de tests de facteur *g* et d'intelligence fluide. Nous verrons plus loin que Raven s'est directement inspiré de Spearman pour concevoir ses épreuves.

Les épreuves de Raven sont bien connues des praticiens et des chercheurs, et ont démontré depuis longtemps leur validité. Une littérature considérable leur est consacrée et elles font preuve d'une remarquable longévité puisqu'elles existent depuis environ 70 ans et que leur popularité chez les praticiens ne se dément pas (elles figurent sans doute parmi les épreuves les plus connues/utilisées au monde). Elles ont en outre inspiré de nombreuses épreuves (voir par exemple le test NNAT ou, plus récemment, le subtest Matrices de la WAIS-III et du WISC-IV).

Il existe trois versions des *Progressive Matrices*.

- Ces versions reposent sur le même type de tâche mais correspondent à trois niveaux distincts de difficulté ;
- La tâche consiste pour le sujet à sélectionner, parmi plusieurs possibilités offertes, l'élément qui vient le mieux compléter une série proposée ;
- Les matrices (doubles séries en lignes et en colonne) comportent quatre éléments (2 lignes et 2 colonnes) ou neuf éléments (3 lignes et 3 colonnes), l'ensemble correspondant au problème à résoudre ;
- La tâche du sujet consiste à découvrir les règles d'organisation (de transformation) de ces différents éléments, puis à appliquer ces règles afin d'identifier la réponse pertinente ;
- La réponse est donnée en choisissant un élément dans un ensemble comportant 6 à 8 réponses possibles.

Un exemple d'item est donné dans la figure 4.1 : le sujet doit sélectionner, parmi les 8 éléments possibles, celui qui vient compléter la série proposée.

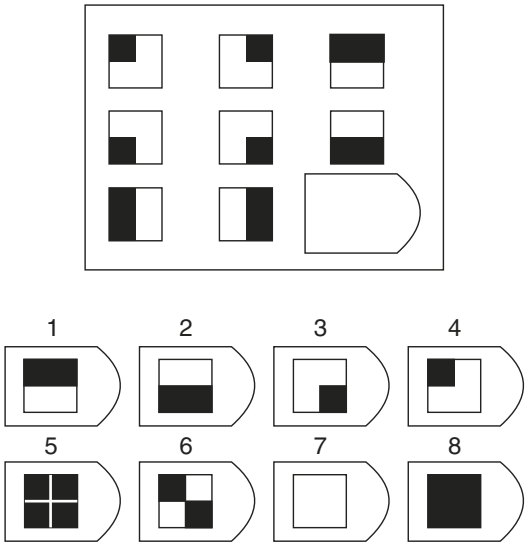

Figure 4.1
Exemple d'un item (fictif) des* Progressives Matrices de Raven *(d'après Huteau, 2002, p. 47).

> ***Exemple***
>
> Dans cette tâche complexe le sujet doit prendre en compte l'ensemble des informations disponibles, et ne pas sélectionner trop vite une solution qui lui apparaîtrait à première vue correcte, mais qui ne comporterait pas en réalité tous les éléments constitutifs de la bonne réponse. L'une des erreurs prototypiques (voir plus loin la partie consacrée à l'analyse des erreurs) est justement la sélection d'un distracteur proche de la bonne réponse, mais ne comportant pas toutes les caractéristiques de celle-ci.

La première version des *Progressives Matrices* date de 1938. C'est la version *Progressives Matrices Standard* ou SPM (ou encore appelée PM38). Cette version *standard* se caractérisait au moment de sa conception par un niveau moyen de difficulté. Ce niveau étant trop facile pour des adultes de haut niveau, il justifia la construction en 1943 d'une version plus difficile : les *Advanced Progressives Matrices* ou A.P.M.

Mais pour les enfants, la version *standard* a été jugée cette fois trop difficile, d'où la diffusion en 1947 d'une version en couleur, assez attrayante pour les enfants : les *Progressives Matrices Couleur*[1]. Au final, on recense

1. Une version encastrable (avec manipulation) destinée à de jeunes enfants (ou enfants présentant un handicap) est également diffusée en France (Kaufman *et al.*, 1993).

donc trois versions des tests Progressives Matrices capables de couvrir une large gamme de niveaux différents et adaptées à un large public allant des enfants aux adultes de haut niveau. Chaque version dispose de son propre manuel et il existe également un manuel commun d'introduction aux tests de Raven (Manuel des Raven, section 1).

Nous ne présenterons ici que les versions utilisables auprès d'adolescents et d'adultes : la version standard SPM et la version APM (diffusées par les ECPA).

Présentation de la version SPM de Raven

Cette version SPM est la version originelle des *Progressives Matrices de Raven* destinée à des sujets de niveaux d'études qui correspondent au collège et au lycée. La première édition date de 1938. Elle comportait 60 items, catégorisés en cinq classes et ordonnés selon leur niveau de difficulté (d'où l'appellation *progressive*). En 1956, de légères modifications, concernant en particulier l'ordre de passation de certains items, ont été adoptées, et la dernière version éditée en France en 1998 est similaire à celle de 1956.

Le manuel commun des Raven (manuel section 1) contient une introduction générale aux épreuves de Raven, et le manuel spécifique à la version SPM (manuel section 3) présente un ensemble de données pour cette version. Des étalonnages complémentaires figurent dans un fascicule séparé, édité en 2003.

➤ Les bases théoriques

Les SPM de Raven ont été créés à partir des hypothèses de Spearman concernant le facteur *g*. En effet, elles sont destinées à évaluer l'*aptitude éductive,* c'est-à-dire la capacité du sujet à percevoir des relations entre différents éléments. Pour Raven,

> « L'aptitude éductive est l'aptitude à créer de nouveaux insights, à percevoir, et à identifier des relations. » (Manuel section 3, p. 5)

Spearman (1927) distinguait *l'éduction de relations et l'éductions de corrélats*. Ces termes correspondent aux formes de raisonnement que nous appelons aujourd'hui pour la première *l'induction* (processus d'extraction d'une règle générale à partir d'exemples particuliers) et pour la seconde *la déduction* (processus d'application d'une règle générale pour générer une réponse particulière). Ces deux formes de raisonnement sont nécessaires pour

résoudre les épreuves de Matrices : il s'agit de repérer les lois de progression entre différents éléments d'une même série et de les appliquer ensuite afin d'identifier l'élément qui vient compléter la série.

Même si les auteurs des manuels font une distinction entre aptitude éductive et facteur *g*, ils indiquent que les matrices donnent bien « l'une des meilleures mesures de *g* » (Manuel section 1, p. 19).

➤ Les items de la version SPM

Les 60 items de cette épreuve sont organisés en cinq séries de 12 items. Rien n'indique cependant au sujet cette organisation, car les items se suivent de façon continue. Quelle que soit la série, la tâche consiste toujours à sélectionner, parmi plusieurs possibilités (6 ou 8 possibilités selon les séries), la figure qui vient compléter l'ensemble.

L'épreuve est organisée selon un ordre de difficulté croissant, aussi bien au sein d'une même série qu'entre les séries. Ainsi, chaque série (de A à E) débute par un item très facile dont la solution « saute aux yeux » (pour reprendre les termes des auteurs : Manuel section 3, p. 5) et se poursuit par des items reposant sur la même logique de résolution mais dont le niveau de difficulté augmente progressivement. L'objectif étant de familiariser le sujet avec le type de raisonnement spécifique à la série d'items afin de lui fournir une possibilité d'apprentissage en cours de tâche. Cet objectif d'apprentissage en cours d'épreuve est d'ailleurs clairement indiqué par les auteurs du manuel (Manuel section 1, p. 65).

C'est aussi par rapport à cet objectif d'apprentissage qu'il faut entendre le terme « *Progressives* » : la résolution des items de chaque série repose sur la capacité à tirer profit de l'expérience acquise dans la résolution des items précédents. Le test évalue alors en partie la capacité du sujet à exploiter l'expérience qu'il vient d'acquérir. Attention cependant : cette dimension progressive de la tâche ne doit pas être confondue avec de réels tests d'apprentissage utilisés dans le cadre d'une procédure d'évaluation dynamique (voir chapitre 7).

Une expérience intéressante, réalisée par Vigneau *et al.* (2001), vient confirmer l'effet positif de la progressivité de la difficulté. Les auteurs ont fait passer les items du test de Raven en ordre inverse de la version habituelle. Les résultats obtenus indiquent que l'épreuve devient alors plus difficile, ce qui peut précisément s'expliquer par l'absence de cette possibilité d'apprentissage en cours d'épreuve.

Comment se différencient ces cinq séries d'items ?

1. **Les problèmes de la série A**, première série des SPM et donc série la plus facile, sont particuliers. Chaque item se présente comme un « dessin » dont il manque une partie. Dans cette série, la tâche consiste donc à compléter le dessin proposé en se représentant les caractéristiques du morceau manquant et en sélectionnant la réponse adéquate parmi les six possibilités de réponse offertes. L'aspect visuel et figuratif du traitement de la tâche est dominant dans cette série.
2. **Les items de la série B** ont la forme d'une matrice à quatre éléments dont celui à trouver (situé en bas à droite) avec six possibilités de réponse. Les premiers items de cette série sont assez simples : il s'agit de répéter les configurations proposées. Le niveau de complexité de la tâche augmente ensuite progressivement et nécessite le repérage des lois de transformation afin de sélectionner la bonne réponse.
3. **À partir de la série C**, les items sont plus difficiles car ils prennent la forme de matrices à neuf éléments (et non plus quatre), et la bonne réponse figure parmi huit propositions et non plus six (voir l'exemple d'item de la figure 4.1). Le nombre d'informations à prendre en compte augmente ainsi sensiblement. La tâche est similaire à celle des items les plus difficiles de la série précédente : il s'agit de repérer les lois de transformation expliquant les transformations en ligne et en colonne.
4. **Dans les items de la série D**, le sujet doit prendre en compte les règles non plus de transformation mais de combinaison des différents éléments du problème de manière à repérer les caractéristiques de la bonne réponse. Par exemple, il doit identifier la règle « chaque ligne et chaque colonne doit comporter le même nombre d'éléments du même type ».
5. D**ans la dernière série du test, la série E,** la tâche est un peu différente car il s'agit d'une logique de type addition ou soustraction de figures.

Au final, le test SPM comporte donc 60 items. Le score brut du sujet, somme des items réussis, peut donc varier de 0 à 60.

➤ Remarques sur le SPM

Avant d'analyser les qualités métriques du SPM, nous voudrions proposer ici deux remarques générales sur cette version, l'une concerne les limites de son utilisation en temps limité, l'autre concerne ses modalités de réponse.

1. Cette organisation en cinq séries présente un inconvénient pour une passation en temps limité. Comme nous l'avons indiqué, les items sont présentés dans un ordre de difficulté croissant au sein de chaque série[1], avec des premiers items très faciles, dont la solution doit « sauter aux yeux », plus faciles que les derniers items de la série précédente. Cela donne au sujet la possibilité, lorsqu'il est confronté aux derniers items d'une série, de ne pas perdre trop de temps à chercher la bonne réponse et de passer directement aux premiers items de la série suivante. Chaque bonne réponse comptant pour un point dans le score total, quel que soit le niveau de difficulté de l'item, deux sujets de même niveau de compétence peuvent obtenir au final des scores différents en fonction de leur style de réponse.

En effet, si l'un adopte cette stratégie de réponse consistant à sauter les items les plus difficiles, alors que l'autre sujet préfère, pour différentes raisons, résoudre chaque item, même s'il doit passer plus de temps sur les plus difficiles, le premier sujet obtiendra probablement un score total plus élevé que le second, en particulier si ce dernier n'a pas eu le temps de traiter tous les items.

Cette possibilité de biais est d'ailleurs évoquée par les auteurs dans le manuel général des Raven. Ils conseillent fortement d'utiliser cette version SPM en temps libre (Manuel section 1, p. 66). Il nous semble que la majorité des utilisateurs du SPM ne connaît pas ce risque potentiel de biais pour une passation du test en temps limité. Nous reconnaissons que le manuel spécifique du SPM (Manuel section 3) ne les y aide pas car la plupart des étalonnages figurant dans cette section sont en temps limité !

Nous avons ici un bon exemple de l'utilité pour le praticien de consulter soigneusement, avant d'utiliser un test, les informations figurant dans le manuel.

1. Les séries étant elles-mêmes présentées selon leur niveau de difficulté (la série A étant la plus facile, la série E la plus difficile).

Recommandation

Nous recommandons donc vivement d'utiliser la version SPM en temps libre, et de préférer, si l'on souhaite effectuer les passations en temps limité, utiliser la version APM (version *Advanced*) car dans cette version, comme nous allons le présenter plus loin, d'une part la phase d'apprentissage est distincte de la phase évaluation et, d'autre part, les items ne sont pas organisés en série ce qui évite ce type de biais.

2. Notre seconde remarque porte sur les modalités de réponse. Nous avons indiqué que dans le test SPM, comme d'ailleurs dans les autres versions des tests de Raven, le sujet ne crée pas la réponse mais il la choisit parmi plusieurs possibilités proposées selon le principe des réponses à choix multiples. Comme nous l'avons indiqué, dans les séries les plus faciles, le nombre de choix possibles est de six (la bonne réponse + cinq distracteurs[1]), puis ce nombre augmente à partir de la série C qui comporte huit possibilités de réponse. Nous savons qu'avec des réponses de type QCM, la possibilité de trouver la bonne réponse par « hasard » n'est jamais négligeable. C'est pour cette raison qu'il est nécessaire, d'une part, de proposer un nombre significatif de distracteurs (ce qui est le cas ici) et, d'autre part, de s'assurer de l'égale attractivité de chaque distracteur. Sur ce dernier point, il nous semble que pour certains items du SPM, les caractéristiques de certains distracteurs sont si éloignées d'une réponse probable qu'un sujet peut rapidement les écarter, ce qui lui laisse au final un nombre plus faible de possibilités de réponses avec une probabilité non négligeable de trouver quelques bonnes réponses par « hasard ».
De plus, le fait de fournir les réponses possibles, lui permet de mettre en œuvre une stratégie consistant à « essayer » chaque matrice en l'appliquant mentalement sur la partie problème, stratégie du type essais/erreurs qui, selon nous, relève probablement d'un autre type de logique que celle qui est visée par cette épreuve. Pour ces raisons il est souvent préférable d'élaborer des épreuves dans lesquelles le sujet doit produire sa réponse, comme c'est le cas par exemple dans les tests de type « dominos » ou séries logiques (voir plus loin en 4.4).

1. Rappelons qu'un distracteur correspond à une possibilité de réponse incorrecte.

➤ Les qualités psychométriques de la version SPM

Depuis la première édition du test SPM en 1938 de très nombreuses études ont été menées qui témoignent de la fiabilité des mesures effectuées par ce test. Les manuels de Raven ne reprennent d'ailleurs qu'une partie de ces différentes études. Cette fiabilité explique sans doute en grande partie la longévité de ce test et son succès.

Le manuel présente les résultats des principales études concernant les qualités métriques des SPM, études menées sur différents pays et sur des échantillons variés (public scolaire, public adulte...). Ces études sont tellement nombreuses que nous ne pourrons en présenter ici une synthèse exhaustive (nous renvoyons le lecteur intéressé vers les manuels). Nous indiquerons cependant quelques résultats qui nous semblent parmi les plus représentatifs.

Les indicateurs de validité

Validité concomitante

Il s'agit ici de vérifier la liaison existante entre ce que mesurent les SPM et ce que mesurent d'autres tests d'intelligence.

Chez les enfants et les adolescents anglophones on observe des liaisons comprises entre .54 et .86 avec des échelles d'intelligence comme celle de Binet ou celle de Wechsler, avec des liaisons plus élevées au niveau du Q.I. Performance qu'au niveau du Q.I. Verbal. Cet aspect n'est pas surprenant car le Q.I. Performance est plus proche de l'intelligence fluide que le Q.I. Verbal. On constate d'ailleurs de façon générale dans de nombreuses études, que les liaisons avec les SPM sont plus élevées avec des tests non verbaux.

Ce constat vaut également pour les comparaisons avec les versions Wechsler pour adultes, comme l'indiquent les données du tableau 4.1.

Tableau 4.1

Corrélations entre les SPM et la WAIS-III (d'après Grégoire, 2004, p. 224).

WAIS-III	SPM
QI Total	.64
QI Verbal	.49
QI Performance	.79
Subtest Matrices	.81

On retrouve bien ici les résultats observés avec les autres versions des échelles de Wechsler : les SPM corrèlent de façon relativement importante avec le QI Total (.64) mais la corrélation est plus élevée avec le QI

Performance (.79) qu'avec l'échelle verbale (.49), et cette corrélation est encore plus élevée avec le subtest Matrice (.81) qui est, rappelons-le, directement inspiré des tests de Raven.

Les liaisons entre les SPM et des évaluations de connaissances sont moins élevées : corrélations variant de .20 à .60 selon les recherches (Manuel section 3, p. 25).

On peut regretter que le manuel ne fournisse pas, sur ces aspects, de données plus complètes relatives à des populations françaises.

Validité prédictive

Les corrélations des SPM avec des critères liés à la réussite scolaire ou à la réussite professionnelle varient largement selon les études. Les valeurs sont justes significatives pour certaines et beaucoup plus élevées (allant jusqu'à .70) pour d'autres (voir Manuel section 3, p. 26 et 27). La plupart des études citées dans le manuel sont assez anciennes mais c'est sur cette base que le test a acquis une bonne réputation de validité prédictive. Les réalités scolaires et professionnelles ayant largement évolué, des études plus récentes sur cet aspect seraient éminemment souhaitables.

Validité de concept

Les SPM, comme les autres tests de Raven sont souvent considérés comme fournissant une mesure relativement pure de facteur *g*. Cette conception est à nuancer. Les analyses factorielles confirment effectivement une forte saturation en facteur *g* des SPM. Mais les études concluent également à une composante spatiale non négligeable. Pour Carroll[1]

> « Les performances aux matrices de Raven sont à la fois déterminées par l'intelligence fluide (niveau II) et par le facteur induction (niveau I). »

Toutefois, Carroll observe qu'au niveau I, le facteur Visualisation « joue un rôle » (Carroll, 1993 cité par Grégoire, 2004, p. 229). Enfin, on peut également observer une relation entre SPM et traitement simultané de l'information (Manuel section 3, p. 29).

Par ailleurs l'hypothèse de la verbalisation (interne) de la démarche de résolution par le sujet, et de son effet sur la performance, ne peut plus être écartée. On peut donc considérer, comme d'ailleurs l'indiquent les auteurs (p. 30), que ce que mesure le test SPM est bien proche du facteur général, et donc de l'intelligence fluide, mais n'est pas pour autant une mesure pure,

1. Voir les propositions de Carroll sur le modèle hiérarchique de l'intelligence dans le chapitre 1.

en raison principalement du rôle probable des aptitudes spatiales et, dans une moindre mesure, verbales.

Les indicateurs de fidélité

La consistance interne

La plupart des études de fidélité interne (méthode *split-half*) concluent à une fidélité élevée avec des coefficients de l'ordre de .90 ce qui est très satisfaisant (Manuel section 3, p. 18).

La fidélité test-retest

Différentes études font état de coefficients autour de .90 pour des intervalles assez courts entre les deux passations et autour de .80 pour des intervalles plus longs, données qui sont très satisfaisantes.

Les auteurs indiquent une tendance à une baisse progressive des scores avec l'âge, surtout à partir de 50 ans, avec par exemple un score moyen de 48 pour les sujets de moins de 30 ans et un score moyen de 29 pour les plus de 50 ans (d'après le tableau SPM1, p. 21, Manuel section 3), mais il faut prendre ces repères avec prudence car les auteurs ne donnent ici aucune indication sur la constitution des échantillons de sujets. Ce phénomène de déclin des performances avec l'âge justifiera la présence, chez les adultes, d'étalonnages par classes d'âges.

La sensibilité

Le test SPM est adapté à un large public, mais un niveau minimum correspondant à des études de collège est nécessaire. En dessous de ce niveau il faut utiliser la version PM Couleur.

Pour les sujets de niveau supérieur au baccalauréat il est préférable d'utiliser la version APM, plus difficile.

Si le praticien hésite entre la version SPM et la version APM il peut établir un diagnostic rapide avec la première série de la version APM, qui contient 12 items, et proposer ensuite la version la plus appropriée au niveau réel du sujet (voir la présentation de la version APM dans les pages suivantes).

Les étalonnages disponibles témoignent d'une bonne sensibilité des SPM.

➤ La standardisation

La passation

Dans la forme classique papier-crayon, le test SPM est un test collectif. La passation nécessite des cahiers de passation, des feuilles de réponse et des stylos. Les auteurs distinguent les consignes pour une passation individuelle

et celles pour une passation collective. Ils estiment à environ 1 heure le temps de passation.

> **Attention !**
>
> Il faut que le psychologue s'assure des caractéristiques des étalonnages qu'il veut utiliser afin de déterminer s'il doit limiter le temps de passation.
>
> En effet, les étalonnages disponibles sont assez hétérogènes de ce point de vue et l'on trouve certains étalonnages de passations en temps libre, condition préférable, mais également en temps limité (là encore la vigilance est de rigueur car, selon les étalonnages, le temps de passation est de 20 ou 30 minutes !).

Les consignes de passation fournies par les auteurs diffèrent en fonction de la modalité de passation : individuelle ou collective. Pour une passation individuelle, le psychologue va se servir des premiers items de la première série (items A1 à A5) comme items d'exemples. L'épreuve proprement dite ne débutant alors qu'à l'item A6. Mais pour une passation collective, il n'y a qu'un seul item exemple, l'item A1, l'épreuve débutant alors à l'item A2.

Ces différences selon les modalités de passation dans le nombre d'items exemples, et donc dans le nombre d'items pris en compte dans la notation, sont un peu surprenantes et peuvent même être sources de biais pour les sujets de faible niveau, susceptibles de ne pas réussir tous les items de la série A. Il est vrai que, pour ce type de sujet, il est préférable d'utiliser la version Couleur.

Les auteurs indiquent l'existence de versions informatisées des SPM (Manuel section 3, p. 41) mais, à notre connaissance, ces versions ne sont pas diffusées en France.

La cotation

Comme généralement dans les tests collectifs, la cotation est aisée. On accorde ici 1 point par bonne réponse. Le score brut du sujet, somme des items réussis, peut donc varier de 0 à 60 points.

Les feuilles de réponse sont auto-scorables, la cotation ne prend donc que quelques minutes.

Le psychologue est invité à calculer également des scores partiels correspondant à chaque série afin d'évaluer la cohérence des résultats. Il dispose alors de normes de références qui indiquent, pour chaque score total,

la décomposition théorique de ce score en cinq scores partiels (voir tableau SPM II, page 59 du manuel section 3).

Par exemple, pour un score brut de 48 points on doit observer la répartition suivante : 12 points sur la série A, 11 points sur la série B, et 9, 10 et 6 sur les séries suivantes. En cas d'écart trop important par rapport à cette répartition théorique (plus ou moins 2 points) le psychologue peut s'interroger sur la validité des résultats. Le manuel donne l'exemple de scores « truqués » par le sujet, dans le sens d'une sous-évaluation de ses performances dans le but d'obtenir une indemnisation (voir Manuel section 3, p. 48).

Au-delà de cette possibilité, cette approche d'analyse des sous-scores nous semble intéressante dans un autre objectif car elle peut permettre de repérer des patrons de réponses atypiques, par exemple un sujet qui réussirait tous les items des séries les plus difficiles (les séries D et E) mais échouerait à certains items des séries plus faciles. Dans ce cas le profil serait alors intéressant à approfondir afin de tenter de comprendre ces échecs étonnants : peut-on les attribuer au niveau de compétence du sujet ? Doit-on envisager d'autres explications, comme par exemple, d'éventuels biais (erreur d'attention...) ?

Ces possibilités d'analyses des sous-scores nous semblent particulièrement pertinentes dans le cas de passation auprès de personnes ne maîtrisant pas, ou maîtrisant mal, la langue du psychologue et/ou auprès de personnes très éloignées, culturellement, des situations de tests, car elles peuvent permettre de vérifier si le sujet a bien compris les tâches proposées dans les différentes séries d'items.

Autre exemple d'utilisation : repérer les sujets qui ont adopté la stratégie de « sauter » les derniers items des séries difficiles. Toutes ces possibilités d'analyse permettant, très probablement, de réduire l'erreur de mesure, c'est-à-dire de rapprocher ce qui est mesuré par l'épreuve (score observé) du niveau réel de compétence du sujet (score vrai).

Le manuel donne également les tables de correspondance afin de convertir un score brut SPM en score brut des autres versions CPM ou APM (voir Manuel section 3, p. 60).

Les étalonnages

Ils figurent dans le manuel ainsi que dans un fascicule « étalonnages supplémentaires » édité en 2003 par les EAP.

Les étalonnages proposés dans le manuel sont nombreux mais il est parfois difficile de trouver un étalonnage pertinent pour un sujet donné. En effet,

beaucoup d'étalonnages concernent des échantillons de sujets étrangers, le plus souvent de langue anglaise, avec des temps de passation qui ne sont pas toujours clairement indiqués... De plus certains étalonnages sont assez anciens (on trouve ainsi, par exemple, des étalonnages de 1979...).

Dans la plupart des cas, les étalonnages prennent la forme de décilages.

Dans le manuel section 3, sur des échantillons Français, nous disposons des étalonnages suivants :

1. Un étalonnage de 1998 sur 670 enfants de 7 ans à 11 ans $\frac{1}{2}$, passation en temps libre (tableau SPM 1) ;
2. Un étalonnage de 1993/96 sur des collégiens et lycéens, en temps limité de 20 minutes, niveaux 6^{e} à 3^{e}, 1re techno, Bac à Bac +2 (étalonnage INETOP : tableau SPM 3) ;
3. Un étalonnage de 1998, sur 708 candidats emploi jeunes, avec séparation selon les niveaux (de CAP à Bac +3), avec des passations en temps limité (20 ou 30 minutes selon les groupes) (tableau SPM 2) ;
4. Un étalonnage de 1992 sur des publics faiblement qualifiés : 160 jeunes de 16 à 25 ans, avec ici un étalonnage normalisé en 9 classes, mais sans précisions sur le temps de passation (tableau SPM 5) ;
5. Un étalonnage de 1987 sur des ouvriers adultes (distingués par classe d'âge), dont on ne nous précise pas le temps de passation (tableau SPM 4) ;
6. Un étalonnage de 1992 sur 246 candidates des écoles des Hôpitaux de Paris, avec un temps de passation de 20 minutes (tableau SPM 6) ;
7. Un étalonnage de 1989 sur 136 ingénieurs, temps de passation 20 minutes (tableau SPM 7), étalonnage qui ne nous semble pas suffisamment discriminatif. Rappelons que pour ce type de public il est préférable d'utiliser la version APM.

Ces différents étalonnages confirment bien le caractère « tout public » de ce test, adapté pour des publics scolaires, des jeunes peu qualifiés mais également pour des populations niveau bac. Mais nous pouvons remarquer l'hétérogénéité de ces étalonnages, et des conditions de passation (temps libre/limité à 20 minutes/limité à 30 minutes).

Pour un test aussi utilisé que le Raven, il manque un étalonnage représentatif de la population française, comparable par exemple aux normes disponibles pour les échelles de Wechsler. On peut regretter que les éditeurs n'aient pas réalisé un tel étalonnage à l'occasion de la rénovation des manuels en 1998.

➤ Les bases de l'interprétation des scores du SPM

Les étalonnages disponibles sont le plus souvent des étalonnages par centilages qui ne comportent que les seuils suivants : 5, 10, 25, 50, 75, 90, 95.

Prenons comme exemple un extrait de l'étalonnage réalisé par l'INETOP (Loarer, 1996) concernant les élèves de collèges (voir tableau 4.2).

Tableau 4.2
Étalonnage INETOP (Loarer, 1996) du test SPM, niveau collège (d'après le manuel SPM, section 3, p. 70).

		Niveaux scolaires		
		6e	5e	4e
Percentiles	95	48	51	53
	90	46	50	51
	75	43	46	48
	50	40	44	45
	25	36	38	42
	10	30	31	36
	5	27	27	33
Moyenne		38,9	42	44,6
Écart type		6,4	7,1	8,1

Comment interpréter les scores ?

Une première étape consiste à situer le niveau de performance du sujet par rapport à la population d'étalonnage. Avec les données de l'étalonnage on connaît la moyenne (avant dernière ligne du tableau) et l'écart type (dernière ligne du tableau) de la distribution des scores au test SPM.

Par exemple, pour les élèves de 6e, la moyenne est de 38,9 points (sur 60) avec un écart-type de 6,4, et l'on sait qu'approximativement 68 % des sujets se situent entre plus ou moins un écart type de la moyenne, donc ici entre 32,5 et 45,3. Un élève de 6e ayant par exemple un score brut de 31 points se situe donc en dessous de la moyenne des élèves de son niveau scolaire (qui est de 38,9 points pour ce niveau scolaire), et légèrement en dessous du groupe moyen que nous venons de définir (qui regroupe 68 % des élèves), et plus précisément juste au-dessus du percentile 10. Autrement dit, seulement un peu plus de 10 % des élèves de son niveau

scolaire ont un score inférieur au sien. Ce n'est pas le signe d'un bon niveau de performance.

Une seconde étape consiste à utiliser la grille de catégorisation proposée dans le manuel. Les auteurs proposent une catégorisation des sujets en cinq classes symétriques, du groupe I « capacités intellectuelles supérieures » au groupe V « déficience intellectuelle » (manuel SPM, section 3, p. 51) :

- *Classe I* : « Capacité intellectuelle supérieure » si le score atteint ou dépasse le centile 95 des sujets de son groupe d'âge. Le sujet se situe alors parmi 5 % les meilleurs ;
- *Classe II* : « Capacité intellectuelle nettement au-dessus de la moyenne » si le score atteint ou dépasse le centile 75. (classe II+ si le score atteint ou dépasse le centile 90) ;
- *Classe III* : « Capacité intellectuelle moyenne » si le score se situe entre les centiles 25 et 75. Cette catégorie regroupe donc 50 % des sujets. On peut éventuellement indiquer classe III+ si le sujet dépasse le centile 50 et classe III- si le score est inférieur à celui ci ;
- *Classe IV* : « Capacité intellectuelle nettement inférieure à la moyenne » si le score se situe au centile 25 ou au dessous (classe IV- si le score se situe au centile 10 ou en dessous).
- *Classe V* : « Déficience intellectuelle » si le score se situe au centile 5 ou au-dessous. Le sujet se situe alors parmi les 5 % les plus faibles.

Il est effectivement préférable, en particulier en raison de l'erreur de mesure, de caractériser le sujet par une « classe » plutôt que par son score précis (*cf.* la notion d'erreur de mesure présentée dans le chapitre 2).

Pour reprendre notre exemple de l'élève de 6e ayant un score brut de 31 points, donc juste au-dessus du centile 10, il se situe alors ici dans la classe IV « Capacité intellectuelle nettement inférieure à la moyenne ».

Les études de cas

Le manuel ne nous propose aucune étude de cas, ce que l'on peut regretter, mais consacre un chapitre au compte rendu des résultats (voir Manuel section 3, p. 51 à 58). Mais celui-ci présente peu d'intérêt lorsque le sujet a passé uniquement le SPM car ce chapitre est plutôt consacré à l'analyse comparée de deux épreuves de Raven (le test SPM et le test de vocabulaire Mill Hill, qui relève plus de l'intelligence cristallisée et du vocabulaire).

Dans cette partie du manuel on trouve également des informations concernant l'analyse des sous-scores, dont nous avons déjà parlé, à partir

des écarts entre chaque sous-score et des références théoriques (voir tableau SPM II, manuel SPM section 3, p. 59). En cas d'écarts de plus ou moins 2 points, les auteurs invitent à questionner la cohérence des résultats du sujet. Sans reprendre nos propositions développées un peu plus haut concernant l'analyse des sous-scores, nous ne pouvons que conseiller également aux praticiens de se pencher attentivement sur le protocole du sujet : analyser son profil de réponse, repérer les items échoués...

Concernant le diagnostic des erreurs, bien que certains chercheurs en proposent un cadre général d'analyse (pour une revue de questions sur ce point voir par exemple Grégoire, 2004, p. 225-229), la pertinence d'une telle analyse semble réduite pour les auteurs du manuel car, selon eux :

> « Le diagnostic des erreurs demanderait la création d'au moins une nouvelle version du test. » (Manuel section 3, p. 56)

Cette absence est également relevée par Grégoire qui précise que :

> « Aucune donnée n'indique en effet qu'il soit possible de différencier les individus en fonction du type d'erreurs commises ou du type de problème où les erreurs sont observées. » (Grégoire, 2004, p. 229)

La version APM des matrices de Raven

Nous venons de présenter, d'une manière assez détaillée, la version SPM la plus connue des tests de Raven. Nous abordons maintenant, mais de façon plus synthétique, la version APM (*Advanced Progressive Matrices*).

➤ Les bases théoriques de la version APM

Cette version *Avancée* repose sur les mêmes principes théoriques que la version *Standard* SPM, avec des items comparables, donc également destinée à évaluer le facteur *g* et l'intelligence fluide. Nous pouvons repérer les spécificités suivantes de cette version APM :

- Elle est plus difficile que la version SPM car elle est adaptée à des sujets de niveau minimum Baccalauréat. La version APM vise à différencier les sujets qui se situeraient dans les meilleurs scores aux tests SPM (les 25 % les plus performants) ;
- Elle comporte deux séries d'épreuves : le Set I, composé de 12 items, est destiné essentiellement à familiariser le sujet avec les situations problèmes qu'il va rencontrer ensuite ; le Set II, qui constitue réellement le test,

composé de 36 items à résoudre en temps libre ou en temps limité (40 minutes) ;

- Les items ne sont pas présentés comme ceux de la version SPM : on ne retrouve pas l'organisation en cinq séries d'items. Les items sont simplement ordonnés selon leur niveau de difficulté.

La version APM est donc bien une évaluation du facteur *g*, utilisable auprès d'étudiants et d'adultes diplômés. Elle fait l'objet d'un manuel spécifique : Manuel des Raven, section 4.

La première version des APM a été élaborée en 1943, puis rénovée en 1947. Cette première version, qui comportait 48 items a été réduite à 36 items en 1962. Depuis cette date aucun item n'a été modifié. La version française la plus récente est datée de 1998.

➤ Les qualités psychométriques

Les études de fiabilité des APM présentées dans le Manuel (section 4), attestent des bonnes qualités de l'épreuve.

Ainsi par exemple, la fidélité, évaluée par la méthode test-retest, varie de .77 à .92 selon les études, et la consistance interne varie quant à elle de .83 à .87. Ces différentes valeurs sont très satisfaisantes.

Les résultats de différentes recherches corrélationnelles sont également présentés dont nous pouvons retirer les éléments suivants :

- Comme la version SPM, cette version APM est fortement liée à des évaluations non verbales de l'intelligence. On relève par exemple une corrélation de .42 avec le QI Verbal de la WAIS et de .55 avec le QI de Performance ;
- Des liaisons significatives avec des indicateurs de réussite scolaire (validité pronostique) sont également observées ;
- De même des liaisons significatives apparaissent avec des critères de réussite professionnelle, même s'il est difficile, comme le soulignent les auteurs, de prédire la performance professionnelle à partir d'une seule variable. La part de variance expliquée peut paraître ici relativement modeste (autour de 10 %) mais elle reste selon les auteurs « notablement supérieure » aux données relevées sur d'autres tests (Manuel section 4, p. 43) ;
- La validité des APM pour l'évaluation du facteur *g* est donc avérée Les auteurs font cependant (pages 25 à 36 du manuel) une mise au

point utile sur la notion de validité en rappelant que d'autres facteurs que le seul résultat à un test d'intelligence comme, par exemple, le niveau de motivation, doivent être pris en compte pour expliquer le niveau de performance d'un sujet et/ou prédire un comportement futur. Ils proposent ainsi leur propre modélisation de l'intelligence (voir en particulier le schéma APM 1, p. 32) dans lequel l'habileté éductive, évaluée par les APM, n'est qu'un des éléments, certes central, des différentes variables de cette modélisation.

Ces études fournissent des éléments attestant d'une bonne fiabilité des APM sont malheureusement plutôt anciennes (certaines datent des années 1970). Il serait donc utile de disposer d'études plus récentes sur des populations françaises

➤ La standardisation

La passation

Dans le format classique papier/crayon le matériel est composé des deux cahiers de passation (Set I et Set II) et de feuilles de réponse (auto-scorables).

Selon la procédure classique, la passation débute avec le Set I (les 12 items de la série 1), dont les deux premiers items servent d'exemples. On indique au sujet que cette première partie est une série d'essais destinée à lui permettre de bien comprendre la méthode de résolution. Il est d'ailleurs possible de donner au sujet cette première série afin qu'il résolve ces items à son domicile. Pour cette série I le sujet dispose soit de 5 minutes (si la passation du Set II, s'effectue ensuite en temps limité), soit de 10 minutes (dans le cas où la passation du Set II, s'effectue ensuite en temps libre).

Lorsque le sujet a terminé le Set I on procède (sans lui) à la correction. Si le sujet a rencontré des difficultés dans cette série, il est alors préférable de poursuivre l'examen avec la version standard des matrices (SPM). Dans le cas inverse, on lui accorde une courte pause puis on lui présente le livret Set II qui contient les 36 items de la seconde série. La passation est alors en temps libre ou en temps limité (40 minutes).

Le manuel indique quelques variantes selon que la passation est individuelle, collective ou encore lorsque le test est administré sans instructions verbales.

Les auteurs indiquent également l'existence de versions informatisées, versions non disponibles en France.

La cotation

Grâce à des feuilles auto-scorables la correction est aisée et ne nécessite que quelques minutes : on accorde 1 point pour chaque item réussi. Le score brut à la série 1 peut donc varier de 0 à 12 points et le score à la série II, score réel des APM, de 0 à 36 points. Contrairement à la version SPM rappelons qu'il n'y a pas ici de possibilité de calcul de sous-scores.

Les étalonnages

Nous pouvons ici renouveler les observations que nous avions formulées à propos des étalonnages de la version SPM :

- Les étalonnages sont nombreux mais hétérogènes (du point de vue de la composition des échantillons comme de celui des conditions de passation : certains sont en temps libres, d'autres ont un temps limité de 30 minutes, d'autres encore de 40 minutes...) ;
- De plus, ils portent sur des populations de différentes nationalités (britanniques, américaines, allemandes, chinoises...) mais aucun étalonnage ne concerne des sujets français. On peut noter que figurent quelques étalonnages concernant la première série d'item (les 12 items du Set I).

La standardisation à grande échelle la plus récente des APM date de 1992 auprès d'une population britannique. Il s'agit d'une passation en temps libre et d'un étalonnage de type percentile avec les mêmes seuils que la version SPM, soit : 5 ; 10 ; 25 ; 50 ; 75 ; 90 ; 95 (tableau APM 13, p. 89). Il s'agit d'un étalonnage par classe d'âge qui distingue 19 catégories (de 12 ans à plus de 70 ans). Les données sont également fournies sous une autre forme permettant de connaître le rang percentile pour chaque score brut observé (tableau APM 14, p. 90).

En lisant attentivement les informations relatives à cet étalonnage, on apprend que si la passation sur les adultes a bien été réalisée lors de la standardisation de 1992, celle des enfants date en fait de 1979. Il aurait donc été plus clair de présenter séparément ces deux étalonnages.

Le praticien français pourra utiliser cet étalonnage qui semble, au vu de comparaisons réalisées sur de petits effectifs, assez proche de ce qu'on pourrait observer sur des sujets français. Il peut également utiliser les étalonnages qui distinguent différentes catégories professionnelles (tableau APM 33, p. 104). Mais il est extrêmement regrettable que nous ne disposions pas, d'une part, d'un réel étalonnage représentatif de la population française, d'autre part, de normes plus récentes. Des étalonnages par niveaux scolaires et par groupes professionnels seraient également très utiles.

Le manuel propose également une table de conversion qui permet d'estimer le score à la version SPM à partir du score obtenu à la version APM, et inversement (voir tableau APM 27, p. 100 et APM 11, p. 58 pour les scores élevés).

Enfin, on dispose également d'un tableau permettant d'estimer un Q.I. à partir du score catégorisé aux matrices de Raven (tableau APM 29, p. 101).

Attention !

D'une part, il faut bien distinguer ce que représente ici l'indicateur Q.I. (qui diffère de ce qu'il représente, par exemple, pour un Q.I. estimé à partir d'une échelle de Wechsler), d'autre part, tout en nous proposant cette conversion des scores le manuel indique qu'il ne faudrait pas l'utiliser en raison de la distribution non gaussienne des données ! (Manuel section 4, p. 101). Deux bonnes raisons donc d'être particulièrement prudent dans l'utilisation de ce tableau.

➤ Les bases de l'interprétation des scores

Comme pour la version SPM, le psychologue commence par repérer le rang percentile où se situe le sujet, puis il utilise la même catégorisation que celle proposée pour la version SPM : de la classe I « capacité intellectuelle supérieure » à la classe V « déficience intellectuelle ».

On peut regretter, comme nous l'avons fait pour la version SPM, qu'il n'y ait aucune étude de cas présentée dans le manuel.

L'analyse des erreurs

Cette version APM a fait l'objet de recherches concernant l'analyse des erreurs. Quatre types d'erreurs ont ainsi été identifiés (Manuel section 4, p. 15-19) :

- *Solution incomplète* : la réponse sélectionnée ne contient que certains aspects de la bonne réponse ;
- *Mode de raisonnement arbitraire* : réponse aléatoire ou relevant d'un principe de résolution non pertinent ;
- *Choix surdéterminé par des éléments intrus* : choix de la solution la plus complexe, qui combinent tous les éléments présents ;
- *Répétitions* : choix d'une des figures déjà présente dans l'espace problème.

D'après les données du manuel, les erreurs de type 1 et 2 sont les plus fréquentes : elles représentent environ 50 % des erreurs, mais il faut signaler que cette répartition fluctue en réalité, selon le niveau des sujets et selon les items. Le tableau APM 2 (p. 17) fournit ainsi la répartition des erreurs les plus fréquentes et le praticien pourra y trouver quelque utilité.

Mais rappelons également que, pour d'autres chercheurs, une analyse des erreurs ne semble pas réellement justifiée (Grégoire, 2004, p. 229).

Il peut également être intéressant d'analyser ces erreurs à partir des traitements cognitifs nécessaires à la résolution des items. Nous pouvons signaler ici que de nombreuses recherches portent sur cette question depuis celle de Hunt en 1974 jusqu'aux travaux plus récents des années 1990-2000. Nous citerons en particulier deux exemples de ces recherches :

- Carpenter, Just et Shell (1990) ont réalisé une analyse cognitive de cette version des matrices et ont identifié cinq règles de résolution. La résolution de chaque item nécessite la découverte d'une ou plusieurs de ces règles ;
- De Shon, Chan, et Weissbein (1995) ont utilisé les verbalisations des sujets pendant la résolution pour caractériser chaque item des APM selon le type de résolution qu'il nécessite : item analytique, item spatial, item mixte. Selon leur analyse il y aurait par exemple 12 items analytiques, 13 items spatiaux, 10 mixtes et 1 inclassable. Nous renvoyons le lecteur intéressé vers l'article qui propose la typologie complète des 36 items des APM.

Conclusions générales sur les tests de Raven (versions SPM et APM)

1. Des tests fiables pour évaluer le facteur *g* et l'intelligence fluide... à condition de respecter les recommandations des auteurs !

Comme nous l'avons indiqué à plusieurs reprises, les tests de Raven ont largement démontré leur validité comme mesures du facteur g et de l'intelligence fluide. Ils présentent de plus l'avantage de la rapidité de la passation, ainsi que de la cotation, et permettent, si le praticien le souhaite, une analyse approfondie du patron de réponse (analyse des sous scores).

Il faut cependant garder à l'esprit que les auteurs du manuel déconseillent l'utilisation de la version SPM en temps limité en raison du risque de biais dans l'évaluation. Pourtant, comme nous l'avons signalé, la majorité des étalonnages utilisables du SPM ont été établis en temps limité...Ce qui apparaît pour le moins paradoxal !

Il est de fait probable que la majorité des utilisateurs du SPM utilise cette épreuve en temps limité. Il est alors, dans ce cas, nécessaire de s'interroger sur la validité du protocole, par exemple, en analysant la répartition des sous-scores par série.

2. Quelle version utiliser : SPM ou APM ? En temps libre ou limité ?

Le premier critère à prendre en compte doit être le niveau du sujet. Le praticien doit identifier à l'avance, en fonction du niveau d'étude du sujet, la version la plus adaptée (SPM ou APM) ainsi que les modalités de passation (temps libre ou limité en fonction des étalonnages qu'il souhaite utiliser). S'il hésite, il peut faire passer le Set I de la version APM et, en fonction des résultats, sélectionner la version la plus appropriée.

Un second critère : l'importance de la possibilité d'apprentissage au cours du test. Les auteurs conseillent d'utiliser la version SPM (en temps libre) pour les sujets peu familiarisés avec le type de situation proposé, en raison du caractère progressif des items qui fournit au sujet des possibilités d'apprentissage en cours de tâche.

3. Des étalonnages insuffisants

Quelle que soit la version nous avons signalé à plusieurs reprises les limites des étalonnages fournis dans les manuels. Il serait nécessaire de pouvoir disposer :

– d'étalonnages plus récents ;
– représentatifs de l'ensemble de la population Française ;
– d'étalonnages par niveaux scolaires ainsi que d'étalonnages par professions.

De plus, les étalonnages en rangs centiles présentent certaines faiblesses par rapport à la discrimination des sujets (voir à ce sujet Grégoire, 2004, p. 223).

Enfin, on peut regretter d'une façon générale l'absence d'études de cas.

2. Le test NNAT (Test d'Aptitude Non Verbal de Nagliéri)

Présentation de l'épreuve

Le NNAT (*Naglieri Non verbal Aptitude Test*) a été élaboré dans les années 1980 par Naglieri. Il s'agit d'une révision et extension d'un autre test de matrice de Naglieri, le « MAT » (Test de Matrice Analogique), test édité

aux États-Unis en 1985 mais jamais adapté en France. Le test MAT est une épreuve de raisonnement non verbal, assez semblable aux matrices de Raven, mais destiné aux enfants âgés de 5 à 17 ans (manuel NNAT, p. 13).

Directement issu du MAT, le NNAT est donc un test de facteur *g* et d'intelligence fluide qui s'inspire largement des épreuves de Raven, comme on peut le constater figure 4.2.

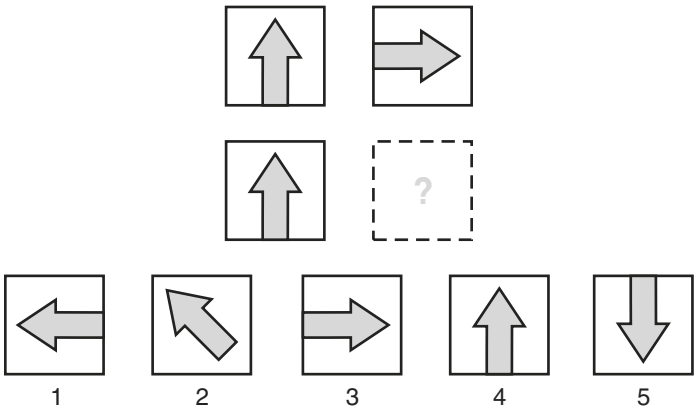

Figure 4.2
Exemple (modifié) d'item du test NNAT (d'après Naglieri, 1998).

Les items sont assez proches de ceux des épreuves de Raven : le sujet doit repérer les règles de progression (de transformation) entre les différents éléments du problème, puis appliquer ces règles afin de sélectionner la réponse correcte.

Le NNAT a été édité en 1996 aux États-Unis, puis adaptée en France en 1998. L'épreuve est éditée par les ECPA et accompagnée d'un manuel de 81 pages.

Le NNAT présente trois caractéristiques principales qui le différencient des autres tests de facteur *g* :

– il se décline en différentes formes,
– il permet le calcul de sous-scores,
– il propose deux types d'étalonnage (par niveau scolaire et par classe d'âge).

➤ Les différentes formes du NNAT

Le NNAT est composé de sept formes (forme A à forme G), chacune adaptée à un niveau scolaire, de la maternelle à la classe de terminale, comme nous pouvons le voir dans le tableau 4.3.

Tableau 4.3

Les différentes formes du test NNAT.

Niveaux scolaires	Maternelle	CP	CE1	CE2-CM1	CM2-6e	5e, 4e, 3e	2e à Terminale
Formes du NNAT	A	B	C	D	E	F	G

Chaque forme est indépendante (matériels et étalonnages séparés) mais toutes comportent 38 items à résoudre en 30 minutes.

➤ Les sous-scores (ou clusters)

Alors que la plupart des tests de facteur *g* (par exemple, les matrices de Raven, les tests de type dominos...) ne caractérisent la performance du sujet que par un score unique[1], le NNAT fournit un score général et quatre sous-scores. Ainsi, à partir de l'analyse des types de raisonnement présents dans les items du MAT, Nagliéri distingue quatre types d'items (appelés également clusters) :

- Complètement de Pattern (P.C),
- Raisonnement Analogique (R.A),
- Raisonnement en série (S.R),
- Représentation Spatiale (S.V).

Chaque item du NNAT étant représentatif d'un type particulier de raisonnement, il est alors possible de calculer, pour chaque sujet, quatre sous-scores. Ces sous-scores apportent une information spécifique sur la performance du sujet dans un type particulier de raisonnement. Il faut signaler, et nous le développerons plus loin, que toutes les formes du NNAT ne contiennent pas obligatoirement ces quatre types d'items.

1. La version SPM de Raven permet, comme nous venons de le voir, de calculer cinq sous-scores mais dans un objectif très différent.

Comment se différencient ces quatre types d'items ?

- Les items de *Complètement de Pattern* (PC) se présentent sous la forme d'un dessin auquel il manque une partie. Le sujet doit « compléter » la figure en sélectionnant le dessin correspondant. Ce sont les items les plus simples, ils sont donc en assez grande quantité dans les formes les moins difficiles du test (par exemple 30 items sur 38 sont de ce type dans la forme A, forme la plus simple du NNAT) et en proportion plus réduite dans les formes les plus difficiles (il n'y a par exemple aucun item de ce type dans la forme G, version la plus difficile du test) ;
- Les items de *Raisonnement Analogique* (R.A) présentent des relations logiques (les règles de transformation) entre les différents éléments du problème. Plusieurs dimensions peuvent varier (forme, hachures, couleur...) et déterminer ainsi le niveau de difficulté de l'item ;
- Les items de *Raisonnement en Série* (S.R) nécessitent pour le sujet de repérer les règles de transformations horizontales et/ou verticales, pour ensuite les appliquer afin de trouver la réponse correcte. Ce type de raisonnement est en fait très proche de celui qui prévaut dans les items de raisonnement analogique ;
- Les items de *Représentation Spatiale* (S.V) nécessitent des opérations d'additions de formes géométriques, de rotations de figures, de pliages... Les auteurs indiquent que ce type d'item est assez difficile. Pour cette raison on les trouve en proportion importante dans les formes les plus difficiles du test (par exemple, 24 items de ce type, sur 38, dans la forme G, version la plus difficile du test NNAT).

➤ La répartition des types d'items dans les tests

Les différentes formes du NNAT comportent le même nombre d'items (38) mais se distinguent dans la répartition des différents types d'items : chaque forme ne comprend pas nécessairement les quatre clusters, et la répartition entre les clusters est différente selon les formes. Les auteurs justifient ce choix en raison des niveaux de difficulté différents de chaque cluster. Par exemple pour la forme A (niveau école maternelle) on observe une surreprésentation des items PC, les plus faciles (30 items soit 79 % des items de la forme A) et une absence des items des catégories les plus difficiles (aucun item S.R ou S.V dans cette forme), et c'est l'inverse pour

la forme G, la forme la plus difficile, qui comporte une majorité d'items SV. La prise en compte de ces combinaisons différentes d'items selon les versions peut permettre alors au psychologue de mieux comprendre ce qui est plus particulièrement évalué dans chaque version du NNAT. Nous y reviendrons.

Contrairement au test SPM de Raven, qui regroupe les items du même type dans cinq séries d'items, dans le NNAT, les différents types d'items sont répartis sur l'ensemble de l'épreuve.

➤ Deux types d'étalonnage

Le NNAT propose deux types d'étalonnages :

– par niveau scolaire (de la classe de Maternelle à la classe de Terminale) ;
– par classe d'âge (de 5 à 15 ans).

Nous détaillerons plus loin l'intérêt de ces étalonnages distincts.

Les qualités psychométriques du NNAT

➤ L'élaboration des sept versions du NNAT

Les items du NNAT sont directement issus de la MAT. Les items ont été sélectionnés de manière à éviter au maximum les influences socioculturelles (manuel, p. 21). De nouveaux items ont été créés pour chaque type de raisonnement de façon à construire les sept formes de test. Les items ont été expérimentés, avec analyse des biais, afin de développer les versions définitives du test.

L'adaptation du NNAT
La version française du NNAT est une adaptation de la version américaine dont elle reprend tous les items. Cette adaptation a été effectuée auprès d'un échantillon de 1 78 élèves de niveaux variés : des classes de grande section de maternelle aux classes de Terminales.

➤ Les indicateurs de la sensibilité

Il s'agit ici de vérifier le pouvoir discriminant du test : permet-il bien de distinguer les sujets ? Autrement dit les résultats sont-ils proches d'une

distribution Normale (de type *Gaussienne*) avec un niveau de difficulté adapté aux populations visées ?

Le tableau 9 du manuel (p. 38-39) fournit les valeurs des moyennes et des écarts types pour chaque forme du test. Sachant que chaque version comporte 38 items, on s'attend à ce que les moyennes se situent autour d'une valeur de 19 points (qui correspond à un pourcentage moyen de réussite de 50 %). De fait, les moyennes varient entre 15,8 à 27 selon les versions, ce qui correspond à des pourcentages de réussite de 41 à 71 % selon les niveaux scolaires (voir tableau 15, p. 48 du manuel), les écarts types variant eux de 5,1 à 8,3.

Si certaines versions semblent un peu trop faciles, ce qui ne facilitera pas la discrimination des sujets de bon niveau (par exemple, on observe un taux de réussite de 71 % pour la version G destinée aux élèves de Terminale), les différentes valeurs témoignent d'un niveau globalement satisfaisant de sensibilité.

➤ Les indicateurs de fidélité

Trois types d'analyse sont présentés. Elles portent sur :

– la fidélité (ou consistance) interne,
– l'erreur de mesure
– la fidélité temporelle.

L'analyse de la consistance interne (rappelons qu'il s'agit ici de vérifier dans quelle mesure tous les items d'un test évaluent bien la même dimension) est réalisée par le calcul des coefficients KR 20 : les coefficients varient ici de .76 à .92 selon les formes (manuel, tableau 8, p. 37). On peut considérer ces valeurs comme satisfaisantes. Par contre, lorsque l'analyse porte cette fois sur les types d'items (les sous-scores ou *clusters*) les variations sont beaucoup plus importantes avec des valeurs comprises entre .23 et .92 en fonction des formes et des *clusters* (voir le détail dans le manuel, tableau 9, p. 38-39). L'interprétation de ces valeurs doit prendre en compte le nombre parfois très faible d'items d'un même cluster mais, et nous y reviendrons, les valeurs les plus faibles peuvent nous questionner sur l'homogénéité de certains *clusters*.

L'erreur de mesure est estimée à environ 2,5 points en note brute[1]. Rappelons que c'est un élément à ne pas oublier dans la phase d'interprétation du score du sujet.

Enfin, troisième type d'analyse, *la fidélité temporelle*. Elle est estimée par la méthode test/retest avec un intervalle de 3 à 5 semaines entre les deux passations. Les coefficients varient de .47 à .82 selon les formes (tableau 10, p. 41). On peut remarquer que la valeur de .47, qui concerne la forme G, semble trop faible pour témoigner réellement d'une bonne qualité de fidélité, mais cette valeur n'est pas commentée dans le manuel. On observe un gain d'environ 3 points entre les deux passations.

➤ Les indicateurs de validité

La validité concourante avec d'autres tests d'intelligence

Une seule étude est présentée ici : elle concerne l'analyse de la relation entre le test NNAT, plus précisément la forme G, et l'épreuve non verbale de la batterie GAT[2], auprès d'un échantillon de 125 élèves. La corrélation obtenue est de .33, ce qui est faible (et les auteurs en conviennent d'ailleurs, *cf.* Manuel, page 42) et beaucoup plus faible que la valeur attendue car les deux épreuves (NNAT et GAT) sont sensées évaluer le même type d'aptitude.

Pour tenter d'expliquer cette (trop) faible valeur les auteurs proposent une argumentation reposant sur le contenu même des items de la forme NNAT utilisée, la forme G, qui est composée d'une grande majorité d'items de visualisation spatiale (24 sur 38) et dont la résolution suppose des manipulations mentales spécifiques : rotation dans l'espace, superpositions de figures complexes (manuel, p. 42). Cette argumentation ne nous semble pas suffisamment convaincante et il aurait été préférable de pouvoir disposer d'autres éléments de validité concourante avec un test comparable, comme par exemple les matrices de Raven, éléments qui viendraient confirmer la validité du test NNAT comme épreuve de raisonnement non verbal de type facteur *g*.

Néanmoins, du fait que le NNAT repose sur le même type de tâche que les matrices de Raven on peut estimer qu'il évalue globalement la même dimension, même si la corrélation observée ici peut nous questionner.

1. Ce qui signifie que « si un sujet obtient une note brute de 24, il a deux chances sur trois que sa "vraie" note se situe entre les notes 22 et 27 » (manuel, p. 37).
2. La batterie GAT (Test d'aptitude générale) évalue l'aptitude au raisonnement logique.

De plus, et c'est étonnant que les auteurs ne rappellent pas ici ce point, les études américaines font état de corrélations élevées (de .63 à .78 selon les formes) entre le NNAT et le MAT (tableau 4, p. 25). Il reste qu'il est regrettable de ne pas disposer, sur un échantillon français, d'autres données que la seule étude présentée. Des études complémentaires mériteraient donc d'être menées.

La validité concourante avec des tests scolaires

Plusieurs études existent dont les résultats principaux figurent dans le tableau 4.4.

Tableau 4.4
Corrélations entre le NNAT et des épreuves scolaires (d'après le manuel, p. 44).

Forme NNAT	Niveau scolaire	Épreuves	Effectif	Corrélation r
D	CE2	NNAT et TNO	50	.45
E	CM2	NNAT et TAS Français	99	.50
		NNAT et TAS Math	97	.53
	6e	NNAT et TAS Français	97	.48
		NNAT et TAS Math.	102	.63
F	5e et 3e	NNAT et TNO	104	.32

Les valeurs se situent autour de .50 et nous pouvons observer les points suivants :

- Les corrélations avec le Test d'Acquisition Scolaire (TAS) varient de .50 à .63, les valeurs étant toujours supérieures avec l'épreuve de mathématiques. La corrélation la plus élevée dans le tableau (.63) étant d'ailleurs observée avec cette discipline ;
- Les corrélations avec le Test de Niveau d'Orthographe (TNO) sont plus faibles, ce qui peut s'expliquer par la spécificité des acquisitions évaluées par le TNO.

Ces valeurs sont conformes à celles généralement observées pour ce type de validité et témoignent d'un niveau satisfaisant de validité du NNAT par rapport à des indicateurs de niveau scolaire.

La validité critérielle avec des appréciations scolaires

Une étude est citée ici qui relie NNAT et appréciations scolaires, à partir d'une échelle en trois points : élève bon, moyen ou faible.

Les appréciations sont générales ou spécifiques à deux matières scolaires (français et mathématiques). L'étude porte sur six niveaux scolaires

(maternelle à CM2) et concerne au total 760 élèves. Des liaisons entre scores au NNAT et appréciations scolaires sont effectivement observées.

La validité théorique

Une première étude porte sur l'aspect développemental de l'épreuve. On s'attend par exemple à observer un pourcentage de réussite plus élevé pour les élèves du niveau supérieur lorsque la même forme s'applique à plusieurs classes. C'est le cas par exemple de la forme F qui s'applique aux élèves de 5e, 4e et 3e : les élèves de 3e obtiennent bien, en moyenne, de meilleurs résultats. Les données confirment ainsi l'aspect développemental du NNAT (voir dans le manuel, tableau 15, p. 48).

Une seconde étude visait à évaluer la validité de la structure du NNAT, structure organisée autour des quatre clusters. À cet effet une analyse factorielle confirmatoire a été effectuée pour chaque forme du NNAT. Les résultats montrent que les *clusters* attendus apparaissent mais, au moins dans certaines formes et pour certains *clusters*, le rattachement de certains items à leur *cluster* théorique de rattachement pose problème en raison de saturations parfois importantes sur d'autres sous-scores que leur sous-score de rattachement.

En effet, lorsqu'on se penche sur les données disponibles dans le manuel (les résultats complets des analyses factorielles figurent en annexe 2 du manuel, p. 75 à 80) on s'aperçoit d'une part, que la saturation observée de certains items avec leur *cluster* (ou sous-score) d'appartenance est faible, voire nulle, d'autre part, que la saturation de certains items avec un autre *cluster* est parfois assez forte. Ces observations, que les auteurs nuancent un peu compte tenu du nombre souvent très faible d'items par *clusters*, peuvent nous questionner sur la typologie des items proposés dans la structure du NNAT. Autrement dit, le calcul de sous-scores à partir des regroupements proposés n'est pas complètement validé par les données. Ce qui explique, d'une part, les conseils des auteurs quand à l'interprétation des sous-scores – ils ne devraient être interprétés que de façon qualitative (Manuel, p. 12) – d'autre part, l'absence d'étalonnages par sous-scores. Nous ne pouvons que souscrire à cette prudence.

Curieusement les auteurs ne nous indiquent pas le décalage entre ce qu'ils ont observé ici, sur un étalonnage français, et les résultats des études américaines qui ont permis de valider la structure du MAT et de définir les *clusters* (Manuel, p. 13).

Ces résultats divergents peuvent-ils s'expliquer par des différences au niveau des outils, par exemple des différences entre les items des deux tests ? Ou sont-ils le signe d'une différence plus profonde, entre les populations,

comme par exemple celle qui a été observé lors de l'adaptation française du WISC-III[1] ? Une étude complémentaire permettrait sans doute d'expliquer ces divergences entre recherches et d'améliorer, du moins on peut l'espérer, la validité des *clusters* proposés dans la version française, donc celle du calcul de ces quatre sous-scores. Nous y reviendrons.

➤ L'étude des biais

Le NNAT est présenté comme étant « culturellement équitable » aussi bien au niveau de l'origine sociale qu'an niveau du sexe des sujets (manuel, p. 1). Quels sont les éléments du manuel qui permettent aux auteurs d'affirmer ces qualités ?

Analyse des différences de réussite selon le sexe

Une analyse comparative selon le sexe a été réalisée pour les différents niveaux scolaires. Les résultats montrent qu'une seule différence est significative : pour les élèves de classes de 2de on observe un résultat supérieur pour les garçons avec une moyenne de 26,55 points contre 23,99 points pour les filles, soit une différence de 2,56 points (manuel, tableau 7, p. 35). Cette seule différence justifierait un étalonnage par sexe pour ce niveau d'étude. Pourtant, cet étalonnage n'a pas été élaboré...

Même si elles ne sont pas statistiquement significatives, les différences de moyennes qui figurent dans le manuel vont quasiment toujours dans le même sens avec une différence en faveur des garçons dans 9 cas sur 12 (voir dans le manuel les données du tableau 7, p. 35). La non significativité des différences est en outre à nuancer compte tenu des effectifs assez restreints des groupes[2].

Ces remarques nous amènent donc à tempérer les conclusions des auteurs concernant l'indépendance de la mesure en fonction du sexe et à attirer l'attention des praticiens, au moins sur le niveau des classes de 2de pour lequel un étalonnage par sexe aurait été nécessaire.

1. Rappelons que dans cette adaptation les auteurs n'ont pas retrouvé sur l'échantillon français l'organisation des aptitudes qui avait été observée sur les sujets américains (voir le chapitre 3 : le WISC-III).

2. Une même valeur de différence entre deux moyennes sera considérée comme statistiquement non significative lorsque les effectifs sont réduits, et significative lorsque les effectifs sont plus importants.

Analyse des différences de réussite selon l'origine sociale

Même si le manuel indique la composition des catégories INSEE du chef de famille (tableaux 5 et 6, p. 29 à 34) aucune donnée n'est fournie concernant la recherche de biais à ce niveau. On ne connaît donc pas les études sur lesquelles reposent les affirmations des auteurs concernant l'indépendance de la mesure par rapport à l'origine sociale.

Au final, on peut observer que les déclarations du manuel concernant autant l'absence de différences entre sexe que l'absence de différences selon l'origine sociale seraient à nuancer, et à compléter par la présentation de données d'études.

La standardisation

➤ La passation

Le test NNAT est une épreuve collective.

Le psychologue doit sélectionner la version correspondant au niveau du (ou des) sujets(s) concerné(s). Rappelons à ce propos qu'il existe sept formes différentes correspondant aux niveaux scolaires suivants :

- Forme A (Grande Section Maternelle),
- Forme B (CP),
- Forme C (CE1),
- Forme D (CE2/CM1),
- Forme E (CM2/6e),
- Forme F (5e, 4e et 3e),
- Forme G (2de, 1re, Terminale).

Pour les versions les plus faciles (formes A, B, C et D) le sujet répond directement sur le cahier de passation ; pour les versions les plus difficiles (formes E, F et G) le sujet répond sur des feuilles de réponse (auto-scorables). À signaler que la première version du manuel (1998) doit être accompagnée d'un additif, additif inclus dans la seconde version du manuel (1999).

Quelle que soit la version utilisée, le cahier de passation comporte 38 items de difficulté croissante. L'utilisation d'items en couleur favorise très probablement l'attrait de l'épreuve, surtout chez les sujets les plus jeunes.

Les consignes précises sont données sur la fiche d'instruction, spécifique à chaque forme. La passation débute par les consignes et les deux items exemples.

Temps de passation : 30 minutes. Avec les consignes et les items-exemples il faut donc compter environ 40 minutes de passation.

➤ La cotation

On attribue classiquement 1 point par bonne réponse, les scores bruts peuvent donc varier de 0 à 38.

Les modalités de cotation diffèrent selon les versions :

- Pour les formes les plus simples, le sujet répond directement sur le cahier de passation, la correction du protocole nécessite une fiche de correction qui donne les bonnes réponses ;
- Pour les formes plus difficiles, avec feuilles auto-scorables, la cotation est plus rapide : il suffit de comptabiliser les croix (choix de la réponse) correctement positionnées.

On obtient ainsi le score total du sujet, mais aussi ses sous-scores (score PC, score RA, score SR et score SV).

➤ Les étalonnages

Ils ont été réalisés en 1997 sur un échantillon de 1 781 élèves de différents niveaux d'études : des classes de Grande Section de Maternelle aux classes Terminales de Lycée. Chaque groupe compte environ 120 à 150 élèves.

Étalonnages du score total

À partir du score brut total on peut distinguer ici deux possibilités :

- Soit le psychologue cherche à situer le sujet par rapport aux élèves de son niveau scolaire : il utilise alors les étalonnages par niveaux scolaires à partir de la note brute totale du sujet (étalonnages normalisés en 11 classes) ; à partir de la note d'échelle (voir plus loin) il est possible d'obtenir le rang percentile du score du sujet par rapport aux différents niveaux scolaires ;
- Soit le psychologue souhaite situer le sujet par rapport aux sujets du même âge, il doit alors transformer sa note brute en une note d'échelle (notes T), puis transformer cette note d'échelle en Index d'Aptitude Non verbale (notes NAI). Au final cet index NAI s'exprime dans une échelle de moyenne 100 et d'écart-type 15, c'est-à-dire dans la même métrique qu'une échelle de Q.I.

Attention !

Cette note NAI, **n'est pas** assimilable à un Q.I. et ne doit pas donc être interprétée comme un Q.I., même si elle en possède la même métrique.

Rappelons que le Q.I, indice d'intelligence globale, est l'indicateur typique des échelles de développement de type échelles de Wechsler, qui reposent sur un ensemble varié de situations et en particulier des items et subtests reposant sur des aptitudes verbales (exemples : le QIV et l'Indice de Compréhension Verbale ICV), ce qui n'est pas le cas ici.

Le NNAT est bien un test non verbal de type facteur *g et non un test de type QI*.

Enfin, le manuel propose également un tableau de conversion entre note d'échelle et « âge équivalent », ce qui est assez surprenant car cette notion, proche de la notion d'âge mental, n'est guère utilisée actuellement.

Étalonnages des sous- scores ?

Nous avons vu que le NNAT permet le calcul de sous scores qui constituent différents indicateurs de la performance. On pouvait alors légitimement s'attendre à disposer d'étalonnages par sous-scores. Cependant, en raison sans doute de la faiblesse de certaines données d'expérimentation (voir plus haut), mais aussi, d'après les indications du manuel (manuel, p. 12), en raison du trop faible nombre d'items qui composent certains clusters, les auteurs n'ont pas élaboré d'étalonnages spécifiques pour chaque sous-score et conseillent de n'interpréter que de façon qualitative ces sous-scores.

Le manuel propose comme seule référence de comparaison les moyennes, écarts types et erreurs de mesure des différents sous-scores pour les différentes formes du test (voir tableau 9 du manuel, p. 38 et 39). Ces éléments nous semblent trop succincts pour être véritablement utiles au praticien.

Nous regrettons cette absence d'étalonnages spécifiques car elle limite l'utilisation de ces indicateurs qui auraient permis de procéder à une évaluation diagnostique. Cela est d'autant plus regrettable que, comme le précisent Bernier et Pietrulewicz :

> « Dans ce type de test, le total ou le score composé est peu significatif ; ce sont les scores partiels ou les pourcentages de bonnes réponses à des regroupements d'items particuliers qui constituent les mesures recherchées. » (Bernier et Pietrulewicz, 1997, p. 224).

Cette possibilité d'analyse diagnostique des résultats du NNAT reste donc actuellement limitée.

Les bases de l'interprétation du ou des scores

➤ Tenir compte de l'erreur de mesure

Avant toute interprétation, il faut se rappeler que la mesure réalisée n'est jamais une mesure fiable à 100 % compte tenu de la notion d'erreur de mesure. Le manuel d'ailleurs nous le rappelle (voir p. 11). Il faut donc, avant toute interprétation, tenir compte de cette erreur (on dispose pour cela des données nécessaires dans le manuel) et, par exemple, entourer le score observé d'un intervalle de confiance. Le manuel donne l'exemple suivant :

> ***Exemple***
>
> Un élève de CM2 qui obtient un score brut de 19 (forme E) est situé dans la classe 6 de l'étalonnage normalisé en 11 classes. Mais sachant que l'erreur type de mesure (Sem) est de 2,6, on peut considérer (avec un risque de 10 %) que le score « vrai » de cet élève se situe entre 16,4 (c'est-à-dire 19 – 2,6) et 21,4 (c'est-à-dire 19 + 2,6), c'est-à-dire entre les notes étalonnées 5 et 7.

Comme dans le cas des échelles de Wechsler, le manuel nous incite à situer le niveau de l'élève non pas à partir d'un score précis, mais à partir d'un intervalle de confiance.

➤ Interpréter le score total

Pour pouvoir interpréter le score total, il faut déjà préciser le type d'étalonnage utilisé : étalonnage par niveau scolaire ou étalonnage par âge. En effet, comme nous allons le détailler, l'interprétation d'un même résultat peut différer en fonction de ce choix.

Lorsque l'élève est « à l'heure », c'est-à-dire scolarisé dans la classe qui correspond théoriquement à son âge (situation d'un élève n'ayant jamais redoublé) les deux normes sont souvent redondantes. Par contre, en cas de retard ou d'avance scolaire, il est pertinent de procéder aux deux possibilités de comparaison.

Prenons un exemple concret :

> ***Exemple***
>
> Un élève de 6^{e} âgé de 13 ans $\frac{1}{2}$, redoublant cette classe et ayant déjà redoublé son CE2, obtient un score brut de 17 points à la forme E du NNAT.

Si on utilise l'étalonnage par niveau scolaire, son score brut de 17 le situe dans la classe 5 de l'étalonnage normalisé en 11 classes (manuel NNAT, tableau 5, p. 71), c'est-à-dire dans la classe centrale. Son score est donc, par rapport aux élèves de 6e, un score moyen.

Mais si cette fois on tient compte de son âge, à partir de son score brut on détermine sa note d'échelle : il obtient une note de 652 (manuel NNAT, tableau 1, p. 55). On convertit ensuite cette note en indice NAI (Index d'Aptitude Non verbale) pour obtenir un indice NAI de 87 (tableau 2, p. 60), c'est-à-dire presque un écart-type en dessous de la moyenne des élèves de son âge.

Ce second type de comparaison, par rapport aux enfants de son âge, montre alors un élève plus en difficulté que lorsqu'on le compare aux élèves de son niveau de scolarisation. C'est sans doute encore plus visible lorsque l'on tient compte de son rang percentile : par rapport aux élèves de 6e, il est situé au percentile 47, c'est-à-dire légèrement en dessous de la médiane des élèves de 6e (46 % des élèves de 6e ont un score inférieur au sien), par contre par rapport aux enfants de son âge (13 ans $\frac{1}{2}$) il est situé cette fois au percentile 19, avec ici seulement 18 % des enfants (de son âge) qui obtiennent un résultat inférieur au sien.

La même illustration pourrait être réalisée, dans un sens différent cette fois, avec un élève ayant une ou deux années d'avance.

Nous voyons bien ici tout l'intérêt que peut présenter cette double possibilité de comparaison pour ces deux types d'élèves.

Pour interpréter le score total il est important, selon nous, car aucun conseil ne figure dans ce sens dans le manuel, de prendre en également en compte la répartition des types d'items dans chaque forme de test.

Comme nous l'avons déjà indiqué, chaque forme ne comprend pas obligatoirement les quatre *clusters*, et la répartition entre les *clusters* est différente selon les formes. La prise en compte de ces combinaisons différentes d'items peut permettre de mieux estimer ce qui est évalué plus particulièrement dans chaque version du NNAT. Ainsi, la version G (niveau Lycée) avec 24 items de type SV, et seulement 7 items SR et 7 items RA, comporte donc une forte majorité d'items qui reposent sur une logique spatiale de résolution (63 % des items de cette forme G sont des items de type SV), tandis que la version D (niveau CE2/CM1), un peu plus équilibrée dans la répartition des différents types d'items (6 items PC, 10 items RA, 8 items SR et 19 items SV) présente de manière moins affirmée cette caractéristique (cette version D ne comportant que 50 % de ces items SV).

Le praticien aura donc intérêt à analyser plus précisément la répartition des items de la version qu'il utilise afin de mieux estimer ce qui est évalué plus spécifiquement dans le test utilisé (la répartition des items est indiquée dans le manuel, tableau 2, p. 3).

➤ Interpréter les sous-scores ?

Nous avons déjà signalé quelques faiblesses psychométriques dans l'affectation de certains items à leur sous-score de rattachement. De plus nous ne disposons pas ici d'étalonnages précis. Ces différents éléments ne peuvent que nous inciter à la prudence dans l'interprétation de ces sous scores.

Nous avons encore ici un bon exemple de l'intérêt, pour le praticien, de lire attentivement les données du manuel afin de mieux cerner les intérêts et les limites du test, et donc les limites des scores et sous-scores qu'il est amené à calculer et à interpréter.

➤ Interpréter le score « équivalent âge » ?

Nous avons déjà indiqué cette possibilité de référence. Mais attention ici, ce score « équivalent âge », qui nous semble proche de la notion d'âge mental, référence qui n'est actuellement plus utilisée, doit être interprétée avec grande prudence et toujours en complément des autres indicateurs étalonnés. Par exemple, il est bien spécifié dans le manuel que ce type de score ne doit pas être utilisé dans l'objectif de décision d'orientation (manuel, p. 17).

➤ L'analyse des erreurs ?

Dans une perspective d'évaluation diagnostique il peut être intéressant de procéder à une analyse des erreurs. Par exemple, en cas d'échec dans un item SR, on peut analyser la (mauvaise) réponse du sujet en recherchant quelles sont la, ou les, règle(s) de transformation qu'il n'a pas prise en compte...

Cette possibilité d'évaluation diagnostique des (mauvaises) réponses ne figure pas dans le manuel mais il nous semble possible d'analyser les erreurs de l'élève afin, par exemple, de faciliter la liaison entre évaluation et remédiation. Cette approche nous semble être directement en lien avec l'un des objectifs affichés de l'épreuve qui est d'identifier les élèves ayant des difficultés d'apprentissage (manuel, p. 5).

Quelques études de ce type existent déjà, comme nous l'avons indiqué, pour les Matrices de Raven, et on pourrait envisager de mener de telles études sur le NNAT.

Conclusion sur le test NNAT

Le test NNAT présente des caractéristiques intéressantes : test collectif, matériel en couleur, attrayant pour les enfants, diversité de formes correspondant à plusieurs niveaux de difficulté, base cognitive d'analyse des items (avec calcul de sous-scores), possibilités de comparaison multiples (étalonnages par niveaux scolaires et étalonnages par âge), étalonnages récents...Mais nous avons présenté certaines de ses limites, en particulier les limites d'interprétation des sous-scores (ou *clusters*).

Ce test NNAT est tout à fait adapté à une utilisation dans un cadre scolaire, il est par exemple utilisé par certains psychologues de l'Éducation Nationale auprès d'élèves de 6e en difficulté scolaire, afin de repérer ceux qui pourraient ensuite faire l'objet d'un examen individuel approfondi.

Enfin, comme nous l'avons développé, une utilisation dans le cadre d'une évaluation diagnostique, avec analyse des profils de réponse, et analyse des erreurs, nous semble intéressante à mener.

3. Les tests D48, D70 et D2000

Présentation des tests

Les tests de type « dominos » sont assez bien connus des psychologues français. Rappelons que dans ces épreuves il s'agit de trouver les deux faces d'un domino qui vient continuer une série proposée. La figure 4.3 nous présente un exemple de ce type d'item.

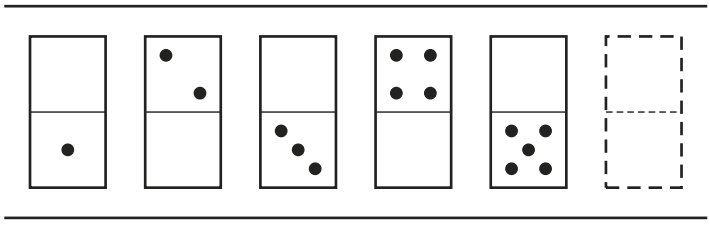

Figure 4.3
Exemple d'item d'un test de type « dominos » (ECPA).

Le sujet doit indiquer les valeurs du domino qui figure en traits pointillés et qui complète la série proposée.

Nous disposons actuellement de trois versions de ce type d'épreuve : les D48, D70 et D2000 (éditées par les ECPA).

La première version française de ce test date de 1948, d'où son nom : le D48. Cette épreuve est directement inspirée de l'épreuve anglaise de Anstey de 1943, le test « *dominoes* », expérimenté à la fin de la guerre dans l'armée britannique. Anstey cherchait à élaborer un test concurrent aux matrices de Raven, diffusées quelques années plus tôt en 1938, et élabore cette épreuve qui présenterait, d'après cet auteur, une saturation en facteur général supérieure à celle des *Progressive Matrices* (Manuel D48, p. 4).

Le D48, adaptation française de l'épreuve de Anstey, est un test collectif, de type papier/crayon, qui comporte 44 items, présentés selon un ordre croissant de difficulté. Ce test a été très utilisé, en particulier dans le recrutement, ce qui explique, en partie[1], la nécessité de procéder à des rénovations régulières. En 1970, une première rénovation a été réalisée avec l'élaboration de la version D70.

Ce test D70 a été directement construit à partir du test D48, dans l'objectif d'élaborer une forme parallèle.

Plus récemment une nouvelle version D2000 a été diffusée. Le test comporte maintenant 40 items, avec un temps de passation réduit à 20 minutes, mais il constitue une version très proche des versions antérieures, avec qui d'ailleurs il partage un certain nombre d'items.

Le D2000 est présenté comme évaluant les mêmes dimensions que les versions précédentes, c'est-à-dire l'intelligence fluide, et plus précisément le raisonnement inductif.

Quelle que soit la version, D48, D70 ou D2000, il s'agit toujours du même type de tâche : le sujet doit trouver la règle de progression, c'est-à-dire définir la (ou les) relation(s) existant(s) entre les faces des différents dominos qui constituent une certaine suite logique, puis appliquer cette (ou ces) règle(s) de progression afin de déterminer les caractéristiques du domino manquant.

Il s'agit bien ici d'une tâche d'éduction de relations et d'éduction de corrélats (ou encore d'induction et de déduction), tâche typique des tests de facteur *g*. Mais par rapport aux autres tests de facteur *g* existants (comme par exemple les Matrices de Raven), les tests de dominos présentent la particularité suivante : le sujet doit construire sa réponse et non pas la

1. Une autre cause de ces rénovations régulières est la nécessité d'établir régulièrement des étalonnages récents (effet Flynn).

sélectionner parmi plusieurs possibilités (d'où une probabilité beaucoup plus faible ici de trouver la bonne réponse par hasard).

En raison du support utilisé, des dominos, ces tests sont souvent considérés comme relevant plus spécifiquement d'une logique de résolution de type numérique. Pourtant, une analyse approfondie des items tend à montrer, et nous détaillerons cet aspect plus loin, que cette considération générale ne reflète qu'imparfaitement ce qui est réellement évalué dans ces tests qui comportent, au moins dans certaines versions, une proportion parfois importante d'items reposant sur une autre logique de résolution, et principalement une logique spatiale (Chartier, 2008a).

Dans la suite de ce chapitre nous nous attacherons à présenter les versions les plus récentes : D70 et D2000.

Les qualités psychométriques

La version D70 a été élaborée à partir de la version D48 (reprise de certains items et création de nouveaux items). De la même manière, la version D2000 a été élaborée à partir de la version D70. Les études montrent que les versions D70 et D2000 sont plus difficiles que la version D48.

Nous pouvons déjà remarquer les dimensions réduites des manuels qui ne comportent qu'un faible nombre de pages (18 pages pour le manuel du D70, 30 pages pour celui du D2000).

Pour le D70, nous trouvons des données concernent la fidélité interne, estimée par la méthode *split-half* (corrélation entre les items pairs et les items impairs), avec une corrélation r de .90, ce qui est très satisfaisant. Au niveau de la validité, le manuel indique les résultats d'une comparaison entre D48 et D70, effectuée sur un échantillon de 250 sujets, dans laquelle on observe une corrélation de .79. Cette valeur témoigne de la proximité des deux épreuves qui évaluent le facteur *g*.

Pour le D2000, le manuel nous indique que des études ont été réalisées sur une première version expérimentale de 60 items, réduite ensuite à 44 items, pour aboutir à la version définitive comportant 40 items. Ces items étant ordonnés selon leur degré de difficulté.

Concernant la sensibilité de l'épreuve, on peut observer que le score brut moyen est de 18,57 points correspond à un niveau de difficulté adapté à l'échantillon de sujets (taux moyen de réussite de 46,4 %). L'écart type de 6,12 points témoigne d'une bonne qualité de la dispersion.

Concernant la fidélité de l'épreuve, le manuel indique une bonne homogénéité interne avec un coefficient alpha de Cronbach de .89. Au niveau des items, les coefficients phi (corrélation items/tests) sont tous significatifs et varient de .10 à .51 (manuel D2000, tableau 2, p. 23).

L'erreur de mesure est estimée à 2,02 points.

Pour ce qui est de la validité, trois recherches contribuent à l'apprécier :

– Une comparaison avec le D48, sur un échantillon de 96 sujets, sur lequel on observe une corrélation de .69 ;
– Une étude comparative avec le R2000 (test de raisonnement et de flexibilité mentale[1]), sur 398 sujets, avec une corrélation de .57. Cette corrélation, plus faible qu'attendue, entre deux tests mesurant l'intelligence fluide est expliquée par l'aspect *flexibilité du raisonnement* qui ne serait présent que dans le R2000 (manuel D2000, p. 24) ;
– Une étude avec un test de *coping* (le CISS), qui conclue à l'absence de liaison avec cette dimension.

Étonnamment, on ne dispose pas d'étude confrontant les mêmes sujets aux versions D70 et D2000.

Les items des tests de dominos

Les items se présentent souvent en ligne comme l'exemple de la figure 4.3 mais d'autres formes de présentation figurent dans les épreuves, comme par exemple des dominos disposés en « étoile ». Quel que soit le type de présentation, la tâche reste la même : identifier les valeurs du domino manquant.

On trouve dans le manuel du D2000 la référence à un article de Dickes et Martin (1998) dans lequel les auteurs distinguent quatre types d'items à partir de l'analyse des items impairs du D70 :

- Les items *spatiaux* : ils nécessitent une stratégie de résolution spatiale.

  « Dans ce cas, le sujet peut s'appuyer sur leur symétrie, répétition, inversion etc. pour trouver la bonne réponse » (Dickes et Martin, 1998, p. 35) ;

- Les items *numériques* : il s'agit ici d'appliquer des règles d'incrémentation entre les faces des dominos pour trouver la bonne réponse (par exemple : ajouter 2 sur une face, retrancher 1 sur l'autre face...)

1. Le test R2000 fait l'objet d'une présentation un peu plus loin.

« L'incrémentation peut se faire sur des faces contiguës ou en alternance. » (Dickes et Martin, 1998, p. 35) ;

- Les items *mixtes :* dans ce cas la résolution d'une des faces est effectuée par une règle spatiale tandis que l'autre face nécessite l'application d'une règle numérique ;
- Les items *arithmétiques* : la solution est trouvée ici par l'application d'une règle arithmétique simple (de type a + b = c) entre les faces de trois dominos.

Les auteurs de cet article nous proposent alors une catégorisation des 22 items impairs de l'épreuve D70 et observent, d'une part, une forte proportion d'items spatiaux (ils représentent 8 items, soit 36 % des 22 items analysés) et, d'autre part, des différences dans le niveau moyen de difficulté : les items spatiaux étant les plus faciles, les arithmétiques les plus difficiles.

Enfin, les mêmes auteurs indiquent que cette caractéristique du D70, avec le caractère composite du score total[1], contribue à la validité du D70 comme test de facteur *g*.

Dans le manuel du D70 il est bien spécifié que les items de ce test, directement inspirés des items du D48, peuvent se différencier sur leur logique de résolution mais ces logiques ne sont pas mentionnées.

Dans le manuel du D2000 il est indiqué qu'une catégorisation des items a été effectuée lors de l'élaboration de l'épreuve, avec dans la version provisoire de 44 items la répartition suivante : 21 items numériques, 9 items spatiaux, 8 de type mixte et 6 numériques[2] (manuel D2000, p. 7). Mais aucun détail supplémentaire n'est donné pour la version définitive comportant 40 items... (voir plus loin nos propositions de catégorisation des items du D2000).

La standardisation

➤ La passation

Les règles de standardisation sont comparables pour les deux versions : il s'agit de tests collectifs de type papier/crayon (cahiers de passation et feuilles de réponse).

1. Qui ne semble pas, pour les auteurs, remettre en cause l'unidimensionnalité de la mesure.
2. Il semble qu'il y ait une erreur dans le manuel qui indique à deux reprises des items numériques... Nous supposons qu'il s'agit ici de 6 items arithmétiques (voir manuel D2000, p. 7).

La passation en temps limité : 25 minutes pour le D70 (44 items), 20 minutes pour le D2000 (40 items).

➤ La cotation

La cotation est simple et rapide : on accorde 1 point par bonne réponse. Il faut signaler que la bonne réponse correspond aux deux faces correctes et qu'il n'y a pas de points, ou de $\frac{1}{2}$ point, si l'une seulement des faces est correcte.

Les scores bruts possibles peuvent donc varier de 0 à 44 points pour le D70, et de 0 à 40 pour le D2000.

➤ Les étalonnages

Les étalonnages disponibles dans les manuels nous renseignent sur le niveau des sujets pour lesquels il est possible d'utiliser ces tests.

Quelle que soit la version, on dispose d'un nombre très limité d'étalonnages.

Pour le D70, un seul étalonnage figure dans le manuel (page 17) : un étalonnage normalisé en 11 classes valable pour « la population adulte à partir de 18 ans et de niveau culturel BEPC » (manuel D70, p. 16). Celui-ci a été établi dans les années 1970, auprès d'un échantillon de 623 adultes, âgés de 18 à 45 ans. Les valeurs caractéristiques (moyennes et écarts type) sont indiquées pour différents niveaux scolaires (du BEPC à supérieur au Bac).

Du fait de l'ancienneté de cet étalonnage, il nous semble peu prudent d'utiliser cet étalonnage (en particulier en raison de l'effet Flynn : voir chapitre 1 de ce livre).

Pour la version D2000 l'échantillon d'étalonnage regroupe 682 sujets, âgés de 18 ans à plus de 55 ans. Il doit dater des années 1999, mais aucune date n'est indiquée. On peut remarquer qu'il présente un déséquilibre au niveau du sexe (avec seulement 40 % d'hommes), au niveau de l'âge (avec 53 % de sujets âgés de 18 à 24 ans), au niveau du diplôme (avec par exemple 29,2 % de niveau d'étude Bac +3 à Bac + 5) ainsi qu'au niveau de la profession exercée.

Ces éléments seront à prendre en considération dans la phase d'interprétation des résultats.

Au final, le manuel du D2000 comporte deux étalonnages en 11 classes (on suppose qu'il s'agit de 11 classes normalisées mais aucune information à ce sujet ne figure dans le manuel) :

- Un étalonnage *général*, sur les 682 sujets de l'échantillon ;
- Un étalonnage *réduit*, sur 398 sujets âgés de 18 à 24 ans, postulants à un concours d'entrée à une école de formation paramédicale (niveau Baccalauréat). Cet échantillon comporte 75,1 % de sujets de niveau Bac, et 24,8 % de sujets de niveau supérieur (voir le détail de cet échantillon en page 24 du manuel D2000).

Attention !

Les auteurs du manuel conseillent d'utiliser cet étalonnage uniquement dans les situations à enjeux comparables (situation de concours...).

En effet, pour cet échantillon (N = 398), ils observent un niveau moyen de réussite plus élevé que sur le reste de l'échantillon d'étalonnage (N = 682).

Du fait de l'absence de différence de réussite entre les hommes et les femmes de l'échantillon, il n'a pas été nécessaire d'établir d'étalonnages séparés par sexe.

On peut regretter ici l'absence d'un étalonnage plus représentatif de l'ensemble de la population française et/ou d'étalonnages spécifiques par âge, niveau d'études et professions.

L'interprétation des scores

Aucune étude de cas ne figure dans les manuels.

Dans les tests de type dominos, l'indicateur de la performance du sujet est un score unique, comme, classiquement, dans tous les autres tests de facteur *g*[1].

Après avoir, éventuellement, calculé l'intervalle de confiance (voir manuel D2000, p. 22), le praticien va situer les performances du sujet dans

1. Excepté pour le test NNAT qui propose, comme nous l'avons déjà indiqué, le calcul de quatre sous-scores.

l'étalonnage le plus approprié. Nous disposons ici d'un nombre limité d'étalonnages, présentés sous la forme d'étalonnages normalisés en 11 classes.

À partir de la classe étalonnée dans laquelle se situe le sujet, il conviendra alors d'estimer plus précisément la position du sujet dans cet étalonnage. On ne peut que regretter l'absence, sur les étalonnages des D70 et D2000, des distributions théoriques qui permettraient au praticien d'effectuer ces estimations. Celui-ci peut utiliser les répartitions théoriques que nous avons présentées dans le chapitre 2 (voir tableau 2.3). Par exemple, si le sujet se situe dans la classe 8, ce score signifie qu'environ 72,5 % des sujets obtiennent un score inférieur au sien, que 11,5 % environ des sujets obtiennent un score comparable, et que seulement 16 % environ des sujets obtiennent un score supérieur.

De plus, l'hétérogénéité de la composition des étalonnages ne facilite pas une comparaison précise des résultats du sujet.

Propositions pour une analyse du profil de réponse

Si le praticien souhaite effectuer une analyse plus fine des réponses du sujet, il peut analyser le patron de réponse (*pattern*) du sujet. Cette approche peut permettre, au psychologue comme au sujet, d'aller au-delà de la simple interprétation du score global, de rechercher la compréhension de la performance réalisée par la prise en compte des items réussis et des items échoués. Elle fournit également des informations supplémentaires permettant d'enrichir la restitution des résultats, de faciliter la compréhension et l'intégration des résultats par le sujet.

Pour procéder à cette analyse, nous pouvons nous référer aux premières recherches de Dickes et Martin (1998) concernant la mise en évidence de quatre types d'items dans ce type d'épreuve (voir plus haut la présentation synthétique des principaux résultats de cet article). Mais rappelons ici que les données ne concernaient que la moitié des items de la version D70. Afin de compléter ces premiers travaux nous avons réalisé une analyse plus large portant cette fois sur l'ensemble du test D70 (les 44 items) ainsi que sur la version D2000.

L'approche que nous proposons ici a fait l'objet de plusieurs communications (voir en particulier Chartier, 2002b) et d'un article de

synthèse (Chartier, 2008a). Cette démarche d'analyse peut être qualifiée de diagnostique au sens de Bernier et Pietrulewicz (1997, p. 224[1]).

Notre approche vise donc, à partir d'une analyse cognitive des items, à proposer le calcul de sous-scores, démarche d'analyse analogue à celle présentée pour le test NNAT (voir la présentation de ce test). Il s'agira ensuite de repérer le type d'items (sous-scores ou *cluster*) pour lesquels le sujet montre, par rapport aux autres sujets comparables, un bon niveau de réussite, ou au contraire, un faible niveau. Cette approche permet alors de repérer les points forts et les points faibles de chaque sujet dans les différentes situations présentes dans le test.

La première étape de cette démarche consiste à effectuer une typologie des items. Pour réaliser cette analyse nous avons repris les règles générales de définition de chaque catégorie d'items proposées par Dickes et Martin. Nous avons ainsi catégorisé tous les items de ces deux versions. Ce sont ces analyses que nous allons maintenant présenter.

La seconde étape, qui reste en partie à réaliser, consisterait à vérifier la validité de ces *clusters* et à élaborer des étalonnages pour chaque sous-score de manière à pouvoir situer le niveau de performance du sujet sur ces indicateurs.

➤ L'analyse des items de la version D70

Cette analyse approfondie des 44 items du D70, nous permet de prolonger, et de confirmer, les analyses de Dickes et Martin : les items du D70 peuvent se différencier sur leur logique de résolution. Effectivement, à partir de la typologie proposée par ces auteurs, nous retrouvons bien quatre catégories d'items : items spatiaux, numériques, arithmétiques et mixtes.

La catégorisation que nous proposons des 44 items du D70 figure dans le tableau 4.5.

Nous retrouvons ici globalement les constats de l'étude de Dickes et Martin : le test D70 comporte une majorité d'items spatiaux, qui représentent plus de 45 % des items de l'épreuve, une proportion beaucoup plus faible d'items numériques (27,3 %) et d'items mixtes (20,3 %), et un très faible nombre d'items arithmétiques (3 sur 44, soit moins de 7 %).

L'épreuve D70 est donc assez déséquilibrée au niveau de la répartition des différents types d'items et elle présente davantage d'items relevant d'une logique spatiale que d'items relevant des autres logiques de résolution, ce

1. Voir dans les pages précédentes, concernant le NNAT, leur définition des tests diagnostiques.

Tableau 4.5
Proposition de catégorisation des 44 items du test D70.

Type d'item	Nombre d'items	Pourcentage d'item	Numéro des items
Spatial	20	45,5 %	1-2-3-4-6-8-11-12-13-14-15-16-17-18-19-23-31-32-41-44.
Numérique	12	27,3 %	5-7-10-21-22-29-34-35-36-39-40-42.
Arithmétique	3	6,7 %	37-38-43.
Mixte	9	20,5 %	9-20-24-25-26-27-28-30-33.
Total	44	100 %	

qui va sans doute à l'encontre des représentations concernant ce test (pour nombre de psychologues le test D70 reposerait essentiellement sur une logique numérique...).

Concernant le niveau de difficulté des types d'items nous retrouvons ici le second constat relevé dans l'article de Dickes et Martin : nous observons en effet, sur un échantillon de 382 adolescents (scolarisés en classe de 3e de collège ou de 2e de Lycée) une difficulté plus faible (en moyenne) pour les items spatiaux et les items mixtes, et une difficulté plus élevée (toujours en moyenne) pour les items numériques et pour les items arithmétiques[1].

➤ Analyse des items de la version D2000

L'analyse que nous venons de présenter concernait les items du test D70 mais la même démarche peut s'appliquer à toute épreuve de domino. Nous avons donc poursuivi notre analyse mais cette fois à partir de la version la plus récente de ce test : le D2000. Nous présenterons ici uniquement les résultats de l'analyse des items car nous ne disposons pas encore de données de passation.

Dans cette version D2000 les auteurs du manuel citent l'article de Dickes et Martin, nous indiquent bien qu'une catégorisation des items a été effectuée lors de l'élaboration de l'épreuve mais ils ne donnent pas le détail de cette catégorisation. En l'absence de ces informations nous avons donc réalisé une analyse des items du D2000 à partir des mêmes règles de catégorisation

1. Les constats sur les items arithmétiques méritent d'être nuancés car, d'une part, ils sont situés en fin d'épreuves et tous les sujets ne les ont pas abordés (du fait de la limite de temps), d'autre part, le sous-score arithmétique repose que sur un faible nombre d'items (3).

que nous avions utilisées dans notre analyse du D70. Les résultats de cette analyse figurent dans le tableau 4.6.

Tableau 4.6

Proposition de catégorisation des 40 items du test D2000 (Chartier, 2008a).

Type d'item	Nombre d'items	Pourcentage	Numéro des items
Spatial	9	22,5 %	2-9-10-12-14-30-32-35-40
Numérique	21	52,5 %	1-3-4-5-7-8-13-16-17-19-21-22-23-24-25-26-27-28-29-31-34
Arithmétique	5	12,5 %	8-33-37-38-39
Mixte	5	12,5 %	6-11-15-20-36
Total	40	100 %	

Le constat global est le suivant : la majorité des items de cette version D2000 relèvent d'une logique numérique (21 soit 52,5 % des items de l'épreuve), une plus faible proportion relève d'une logique spatiale (9 items, soit 22,5 %), et une proportion encore plus faible (12,5 %) pour les deux autres logiques de résolution.

➤ Conclusion sur l'analyse des items du D70 et du D2000

Nos résultats confirment donc les premières analyses de Dickes et Martin concernant l'existence de différentes logiques de résolution dans les tests de type domino : nous avons ainsi distingué des items spatiaux, numériques, arithmétiques et mixtes. Les tests domino ne reposent donc pas exclusivement, comme le pense sans doute souvent un certain nombre de praticiens (et de chercheurs), sur un seul type de logique. L'observation d'une pluralité de logique de résolution au sein d'un test de facteur *g* n'est d'ailleurs pas rare, rappelons par exemple les nombreuses analyses sur les Matrices de Raven depuis celles de Hunt dans les années 1970 (Hunt, 1974), jusqu'aux analyses plus récentes, comme par exemple celles de Carpenter, Just et Shell (1990) ou de De Shon *et al.* (1995). C'est d'ailleurs sans doute en raison d'une pluralité de logiques de résolution que ces tests de dominos semblent être de bons représentants du facteur g, comme l'avaient déjà souligné Dickes et Martin (1998).

L'analyse des items des versions D70 et D2000 que nous avons réalisée nous conduit à proposer les commentaires suivants :

- La composition de la version D2000 est sensiblement différente de celle de la version D70 : si le test D70 repose surtout sur des items nécessitant des règles spatiales de résolution, la version D2000 repose plutôt sur des items qui nécessitent des règles numériques. Bien que l'on considère ces deux versions comme proches, cette différence mérite notre attention ;
- Pourquoi ne pas avoir profité de cette rénovation du test des dominos pour équilibrer les différentes catégories d'items ? Ce qui aurait éventuellement permis le calcul de quatre sous scores reposant chacun sur un nombre suffisant d'items ;
- On peut s'interroger sur la pertinence de placer la majorité des items Arithmétiques en fin d'épreuve du D2000 (les items 37, 38 et 39 soit 3 items sur 5) ce qui a comme conséquence de limiter grandement le nombre de sujets qui auront l'occasion d'aborder ces items, par manque de temps.

Recommandation

Ces deux séries d'observation, d'une part la diversité des logiques de résolution des items, d'autre part, le déséquilibre dans leur répartition selon les versions des tests (D70 et D2000), peuvent être des informations utiles au praticien.

En effet, elles lui permettent de mieux connaître ce qui est principalement évalué par chaque version du test et concourent alors à améliorer les données concernant la validité de ces tests. Elles permettent également d'identifier les différentes logiques de résolution qui doivent être appliquées par le sujet tout au long de l'épreuve et contribuent ainsi à l'identification des difficultés rencontrées par un sujet dans un item particulier, ou dans une catégorie d'items.

Plus globalement, la démarche d'analyse des réponses que nous proposons ici permet également au praticien d'enrichir la phase de restitution des résultats qui reste centrée trop souvent, sur ce type de test, autour de l'interprétation du seul score total. En effet, et ceci est valable plus généralement dans tout test de facteur *g*, il est souvent difficile de dépasser le constat du seul score total car on sait que ce type d'indicateur donne peu d'information sur les conditions de réalisation de la performance (Huteau et Lautrey, 1999a ; Huteau, 2001 ; Lautrey, 2001). Cette démarche d'analyse que nous proposons ici, qui se place plus globalement dans le cadre d'une

évaluation diagnostique, permet de fournir quelques pistes explicatives concernant la performance du sujet dans l'épreuve. Pistes qu'il est possible ensuite de développer avec le sujet, par exemple, lors d'un entretien de restitution.

Cette approche diagnostique rejoint également les préoccupations actuelles de chercheurs qui visent à combiner une évaluation quantitative à une évaluation qualitative dans une approche intégrative de l'intelligence (Rozencwajg, 2005).

Enfin, nous pouvons signaler une autre direction de recherche s'appuyant sur des tests de type dominos. Il s'agit ici de procédures expérimentales, développées par Rémy (2001) et Rémy et Gilles (1999) visant à mettre en évidence des différences interindividuelles dans les stratégies de résolution d'items de type dominos. Dans leur épreuve expérimentale les auteurs ont élaboré des items pouvant être résolus soit par une stratégie spatiale, soit par une stratégie numérique. Et selon la stratégie employée (numérique ou spatiale), la bonne réponse est différente[1]. Ainsi, en analysant la réponse donnée par le sujet à un item, à condition bien entendu qu'il s'agisse d'une des deux bonnes réponses prévues, on peut en inférer directement la stratégie utilisée par le sujet. Ce dispositif permet alors de repérer la stratégie préférentielle du sujet mais également son niveau de flexibilité (utilisation des deux stratégies en fonction des caractéristiques des items). Mais ces recherches ne semblent pas avoir été finalisées par l'élaboration d'une épreuve éditée et/ou utilisable par un praticien.

Conclusion sur les tests de dominos

Les versions D70 et D2000 que nous venons de présenter sont considérées comme des tests mesurant le facteur *g*. Pourtant, sans remettre en question cette considération, les données d'études disponibles dans les manuels nous semblent insuffisants. Il manque, par exemple, une étude de validité portant sur les liaisons entre le D2000 et un autre test de facteur *g* (la faiblesse de la seule étude présentée dans le manuel[2] est d'ailleurs soulignée par les auteurs : voir page 24 du manuel D2000).

1. Il y a donc ici 2 bonnes réponses par item.
2. Rappelons que cette étude porte sur le test R2000, avec une corrélation observée entre les deux épreuves de .57.

De plus, et nous l'avons déjà indiqué, les étalonnages disponibles sont en nombre trop limité : il serait nécessaire de disposer d'étalonnages par âge et/ou par niveau scolaire et/ou par profession.

Il reste que ces épreuves présentent les avantages des autres épreuves de type facteur *g* (rapidité de la passation et de la correction) avec ici un avantage particulier : il est demandé au sujet de créer sa propre réponse, alors que dans les tests comparables il doit uniquement, le plus souvent, sélectionner l'une des possibilités de réponse (épreuve de type QCM). Cette particularité des tests de dominos permet ainsi de réduire grandement le risque de donner une bonne réponse par « hasard ».

Nous avons illustré que ce type de test pourrait permettre une analyse plus fine des réponses des sujets. Pourquoi ne pas avoir profité de cette rénovation pour proposer au psychologue les outils (catégorisation de chaque item, procédure de calcul des sous-scores, étalonnages spécifiques...) permettant d'effectuer une telle analyse ? Une application très concrète des résultats de recherches était donc possible ici et, sans remettre en cause l'intérêt du D2000, on ne peut que regretter le caractère trop classique de cette rénovation.

Nous avons présenté nos propositions de catégorisation des items du D70 et du D2000, avec les premiers éléments d'un cadre d'évaluation diagnostique des réponses des sujets dans ce type d'épreuve.

4. Le test R85/R2000

Présentation de l'épreuve

Le test R2000 (test de Raisonnement, version 2000), édité en 2000 aux ECPA, est directement issu du test R85 (de 1985) et du test de raisonnement de Pierre Rennes de 1952. C'est une épreuve présentée comme étant une mesure de l'intelligence fluide, utilisable chez des sujets de niveaux d'études supérieures.

La particularité de cette épreuve est qu'elle repose sur un matériel assez varié (verbal, numérique et mixte) et évalue alors également la flexibilité du raisonnement, définie ici comme la capacité de passer d'un type de raisonnement à un autre (manuel R2000, page 1).

Les tâches proposées sont donc assez diverses, tant au niveau des supports, qu'au niveau de la tâche réelle qu'il s'agit souvent de découvrir. Il peut s'agir,

par exemple, de continuer une suite logique de chiffres, ou de lettres, ou encore de repérer un intrus...

Les exemples suivants permettent de se faire une idée de la diversité des items :

36 12 24 4 ?

Il s'agit ici de trouver le chiffre qui vient remplacer le point d'interrogation.

chapeau soulier robe miroir gant

Il s'agit ici de comprendre qu'il faut repérer (souligner) l'intrus.

Nous nous centrerons ici sur la présentation de la version la plus récente de ces épreuves : le R2000. Cette version comporte 40 items à résoudre en temps limité (20 minutes).

Les qualités psychométriques

Pour élaborer la forme R2000 les auteurs sont partis de la forme R85 (qui comporte 40 items) et ont élaboré 80 nouveaux items. Ces 120 items ont été testés (à partir de 2 versions parallèles) et 40 items ont été sélectionnés pour la version définitive du R2000 : 15 items verbaux, 10 items mixtes et 15 items numériques.

Cette version définitive a été expérimentée sur un échantillon de 625 sujets adultes, de niveau minimum Bac, en situation professionnelle d'évaluation (recrutement, bilan, gestion de carrière...).

➤ La sensibilité de l'épreuve

Avec un score brut moyen de 15,12 points, soit un taux moyen de réussite de 37,8 %, nous pouvons constater la difficulté de l'épreuve. Cette difficulté est progressive avec 92,8 % de réussite sur l'item 1 et 9,8 % sur le dernier item (voir table 3, p. 23 du manuel). La version R2000 est plus difficile que la version R85.

L'écart type de 6,56 points témoigne d'un bon niveau de dispersion.

➤ La fidélité

L'homogénéité interne est évaluée par le calcul de l'*alpha* de Cronbach : la valeur observée de .89 est satisfaisante.

Les corrélations items tests *phi* sont toutes significatives au seuil de .01, et varient de .06 à .47.

L'erreur-type de mesure est estimée à 2,13 points.

➤ La validité

Concernant la validité, le manuel fait état des résultats de trois recherches :

- Une comparaison avec le R85, sur un échantillon de 62 sujets, sur lequel on observe une corrélation de .67. Le manuel indique un « degré de liaison important » entre ces deux épreuves (manuel R2000, p. 26) mais nous pourrions nous attendre à observer une valeur supérieure ;
- Une étude comparative avec le D2000[1], portant sur 398 sujets, avec une corrélation de .57 (il s'agit visiblement de la même étude que celle qui a été présentée dans le manuel du D2000). La valeur de cette corrélation entre deux tests mesurant l'intelligence fluide, visiblement un peu faible selon les auteurs du manuel, est expliquée, d'une part, par l'aspect *flexibilité du raisonnement* qui ne serait présent que dans le R2000, d'autre part, par la différence de supports (manuel R2000, p. 25) ;
- Une étude avec un test de *coping* (le CISS), qui conclue à l'absence de liaison avec cette dimension.

La standardisation

➤ La passation

Le R2000 est un test de type papier/crayon, dans lequel le sujet répond directement sur le cahier de passation (1 feuille A4, plié en A5). Après les 6 exemples, la passation des 40 items de l'épreuve se déroule en temps limité (20 minutes).

➤ La cotation

La correction est rapide et s'effectue à l'aide d'une grille. On accorde 1 point par bonne réponse. Le score brut peut donc varier de 0 à 40 points.

1. Le test D2000 a été présenté plus haut.

➤ Les étalonnages

L'échantillon d'étalonnage comporte 625 sujets adultes, âgés de 18 à 54 ans. Cet étalonnage doit dater des années 1999 mais aucune date n'est indiquée. L'échantillon comporte des déséquilibres par rapport au sexe, avec une majorité de femme (elles représentent près de 72 % de l'échantillon), par rapport à l'âge, avec une majorité de sujets dans la classe 18-24 ans, ainsi que par rapport au niveau scolaire. Mais seule la différence entre les sexes est significative, avec des résultats en faveur des hommes. Pour cette raison les auteurs proposent un étalonnage séparé par sexe.

Au total quatre étalonnages figurent dans le manuel. Il s'agit d'étalonnages en 11 classes, qu'on suppose être des étalonnages normalisés (mais aucune indication à ce sujet ne figure dans le manuel) :

– un étalonnage *global*, sur les 625 sujets de l'échantillon ;
– un étalonnage *hommes* (sur 175 sujets) ;
– un étalonnage *femmes* (sur 448 sujets) ;
– un étalonnage *réduit*, sur 398 sujets, candidats à un concours.

On peut raisonnablement supposer que cet échantillon de 398 sujets est identique au sous-échantillon de 398 sujets cités dans le manuel du D2000. Mais alors que dans le D2000 il est question de 398 jeunes inscrits à un concours d'entrée dans une école à des formations paramédicales (manuel D2000, p. 24) il est ici question de « jeunes filles inscrites à un concours » (manuel R2000, p. 25). Comme pour le test D2000, les résultats moyens observés sur cet échantillon sont supérieurs à ceux de l'échantillon total, ce qui peut sans doute s'expliquer par la nature de la situation (concours). Les auteurs du manuel proposent également de réserver l'utilisation de cet étalonnage pour des femmes, jeunes, de niveau d'étude Bac et dans des situations à enjeux (concours, recrutement...) (manuel R2000, p. 26).

Comme pour le test D2000, on ne peut que regretter l'absence d'un étalonnage plus représentatif de l'ensemble de la population française et/ou d'étalonnages spécifiques par âges, niveau d'études et professions.

L'interprétation des scores

Aucune étude de cas ne figure dans le manuel (qui ne comporte que 31 pages). L'interprétation des scores suivra ici la même démarche que celle proposée pour le D2000 : il s'agira de situer précisément le niveau de performance du

sujet dans l'étalonnage le plus approprié (nous ne reprenons pas ici l'exposé de cette démarche et renvoyons le lecteur vers la partie interprétation du D2000).

Dans l'interprétation de ce score il faudra bien entendu prendre en compte les spécificités de l'échantillon d'étalonnage, qui sert de référence.

Conclusion sur le test R2000

Ce test R2000 est un test qui semble difficile et qu'il faut réserver aux sujets de niveau d'étude minimum Bac/Bac +2. Du fait de son niveau de difficulté, et du support varié, il peut susciter un niveau élevé de stress lors de la passation. Stress qu'il faudra éventuellement prendre en compte, par exemple en questionnant le sujet dans la phase de restitution des résultats.

Ce test présente cependant l'avantage de discriminer les sujets de haut niveau de qualification (par exemple des ingénieurs). Il est rapide et facile à corriger.

Il mériterait cependant d'être accompagné d'étalonnages spécifiques par niveau d'études et/ou professions.

5. Quelques autres tests de facteur g

Comme nous l'avons déjà indiqué, il n'est pas possible dans cet ouvrage de faire figurer une analyse détaillée de chaque test disponible en France. Nous ne donnerons donc ici que quelques informations sur trois autres tests de même type :

- le test de Culture Fair de Cattell,
- le BLS4,
- le B53,
- le RCC.

Le test Culture Fair de Cattell

Le *Culture Fair Intelligence test* de Cattell, élaboré en 1940, est une des tentatives de mesure de l'intelligence fluide, indépendante de la culture (*culture free*) ou encore culturellement équitable (*culture fair)*. L'une des

spécificités de cette épreuve est de présenter quatre formats d'items afin, justement, d'éviter de désavantager certains sujets par la présentation d'un seul type d'item :

– des complèments de séries,
– des classifications,
– une épreuve de matrice,
– une épreuve spatiale.

Une version de 1986 est éditée par les ECPA mais, selon Grégoire, certaines qualités psychométriques semblent un peu faibles (Grégoire, 2004, p. 236).

Le BLS 4

Il s'agit d'un test assez ancien de Bonnardel, élaboré dans les années 1950 et qui a été rénové en 2000 et diffusé par les EAP. Bonnardel présente son épreuve comme une épreuve de facteur *g* et de potentiel intellectuel (Thiébaut, 2000). Dans cette épreuve le sujet doit continuer une série proposée.

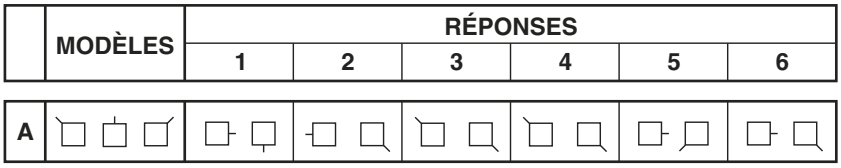

Figure 4.4
Exemple d'item de BLS 4.

Exemple

Dans cet exemple, le sujet doit sélectionner la réponse (parmi 6 possibilités) qui vient continuer le « modèle ». Le BLS4 comporte 30 items de ce type.

On peut signaler ici deux spécificités de cette épreuve :

- Il en existe deux versions : l'une sous la forme de questions fermées (réponse à sélectionner, comme l'exemple de la figure 4.4), l'autre sous la forme de questions ouvertes (réponses à construire). Chaque version possède ses propres étalonnages ;

- Le manuel propose une analyse des erreurs qui permet au praticien d'approfondir les réponses du sujet.

Cette épreuve est adaptée à des sujets de niveau Bac et post-bac. Avec un temps de passation de 10 minutes cette épreuve est assez courte. Ce test est utilisable en procédure d'orientation et en recrutement (Thiébaut, 2000).

Le test B53

Cette épreuve, également de Bonnardel, a été rénovée en 2000 (et diffusé également par les EAP) Il s'agit également ici d'une tâche typique de test de facteur *g* basé sur la découverte de lois de progression entre différents éléments.

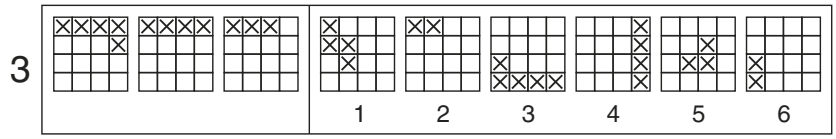

Figure 4.5
Exemple d'item du B53.

Exemple

Le sujet doit indiquer ici quelle est la figure de droite, parmi les six possibilités, qui doit continuer la série proposée.
Le B53 comporte 65 items de ce type (dont les 5 exemples), de différents niveaux de difficulté. Le temps de passation est limité (15 minutes).
La feuille de réponse, auto-corrective, permet une correction très rapide.

Ce test est utilisable auprès d'un public varié : du niveau BEP au niveau Bac + 2 : 10 étalonnages sont disponibles Enfin, on peut signaler que le manuel est commun aux deux tests BLS4 et B53.

Le test RCC[1]

Il s'agit d'une épreuve de raisonnement sur support de cartes à jouer. Une suite de cartes est présentée au sujet qui doit déterminer les caractéristiques

1. Raisonnement sur Cartes de Chartier (Chartier, 2008b).

de la carte qui vient continuer (ou compléter) cette série. La version expérimentale de cette épreuve est en phase d'édition (chez Eurotests). Elle permet de recueillir plusieurs indicateurs (Chartier, 2008b) :

- un score total ;
- deux sous scores : numérique et spatial (en fonction des logiques de raisonnement identifiées) ;
- une analyse des erreurs.

CHAPITRE
5

Les batteries factorielles

Sommaire

LES épreuves que nous allons maintenant présenter sont directement issues des propositions de Thurstone concernant l'existence d'aptitudes différenciées (voir chapitre 1). Elles prennent souvent le nom de *batteries factorielles* car chaque batterie est constituée d'un ensemble de tests. L'une des principales caractéristiques de ce type d'épreuve est qu'elles offrent la possibilité d'établir un profil des résultats du sujet en fonction des aptitudes évaluées. Il s'agit là d'une différence importante avec les tests de facteur *g* qui ne fournissent généralement qu'un score unique.

Les batteries factorielles sont le plus souvent des épreuves collectives, de type papier-crayon. Chaque épreuve d'une batterie vise à évaluer une aptitude définie et fait l'objet de consignes et d'étalonnages spécifiques. Ainsi, en fonction de ses objectifs, le psychologue peut choisir de faire passer la batterie de tests en totalité ou de sélectionner certaines épreuves, ce qui permet une certaine souplesse d'utilisation.

Nous présenterons dans ce chapitre les principales batteries utilisées en France : les batteries NV5, NV7 ainsi que la DAT5.

Chaque épreuve possède ses spécificités, comme par exemple le nombre d'aptitudes évaluées ou les étalonnages disponibles. Par exemple, la DAT5 et la NV7 se distinguent sur le nombre d'aptitudes prises en compte (8 pour la DAT5 et 10 la NV7) mais également sur le public visé, la NV7 étant élaborée pour des jeunes adultes de faible niveau de qualification, alors que la DAT5 est étalonnée sur des publics scolaires de niveau fin de collège et de Lycée ainsi que sur des adultes.

1. La batterie NV7

Présentation de l'épreuve

La batterie NV7 est une création française (Bernaud, Priou, Simonet) éditée en 1993 aux EAP.

L'objectif des auteurs était de créer une batterie multifactorielle d'évaluation des aptitudes destinée à un public faiblement qualifié. Pour élaborer cette épreuve, les auteurs ont sélectionné un certain nombre de tests, diffusés antérieurement, pour les réunir sous forme d'une batterie. La NV7 a été éditée en 1993 mais elle regroupe en réalité des tests beaucoup plus anciens,

dont certains, nous le verrons plus loin, présentent des caractéristiques historiquement marquées (utilisation d'images en noir et banc, style de graphisme des images des items...).

Les consignes et les contenus des items sont adaptés à un public de faible niveau scolaire, sans qualification, ou avec un niveau inférieur au Bac professionnel (voir les étalonnages disponibles).

C'est une épreuve très utilisée actuellement dans les pratiques de bilan de compétences auprès de publics peu qualifiés, mais également dans des évaluations de type « retour à l'emploi » auprès de sujets qui possédaient un niveau de qualification supérieur mais qui, suite à diverses circonstances (accidents, arrêt prolongé de l'activité professionnelle...) s'interrogent sur leur niveau actuel de performance.

Les auteurs ont par la suite élaboré une version plus difficile, la NV5-R, adaptée à des sujets de niveau supérieur (niveau Bac et plus) que nous présentons plus loin.

La batterie NV7 comprend dix épreuves (ou subtests) :

1. Raisonnement déductif (R1),
2. Raisonnement inductif (R2),
3. Raisonnement analogique (R3),
4. Raisonnement pratique-technique (R4),
5. Spatial,
6. Problèmes,
7. Opérations,
8. Attention,
9. Orthographe,
10. Compréhension verbale.

Certaines de ces épreuves sont proches des aptitudes mentales primaires de Thurstone (exemple : les épreuves de raisonnement et d'aptitude spatiale) tandis que d'autres renvoient à des apprentissages scolaires (exemple : Orthographe).

La passation complète de la batterie nécessite environ 1 heure 45 minutes.

Détaillons maintenant chacun de ces tests.

➤ Épreuve de Raisonnement déductif (R1)

Elle évalue la capacité à raisonner du général au particulier et comporte 24 items (dont 2 items d'exemple) à résoudre en 8 minutes. Les items prennent la forme de quatre images ordonnées. Le sujet doit indiquer si

la suite chronologique est respectée (réponse « exacte ») ou non (réponse « inexacte »).

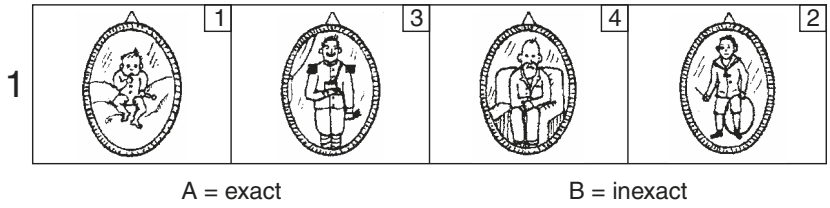

Figure 5.1
Exemple d'item du subtest Raisonnement déductif (R1).

Deux remarques sur cette épreuve :

- Les items qui la composent ont une apparence aujourd'hui « vieillotte » (type de graphisme, images en noir et blanc) ce qui peut avoir un effet sur la motivation du sujet, d'autant plus qu'il s'agit de la première épreuve de la batterie ;
- Le mode de réponse proposé, choix entre la réponse A et la réponse B, a comme inconvénient majeur de laisser une probabilité importante de trouver la bonne réponse par le fait du hasard (50 %). Cela aurait pu être évité en demandant au sujet de produire lui-même le classement des images.

➤ Épreuve de Raisonnement inductif (R2)

À l'inverse de l'épreuve précédente, il s'agit ici de raisonner du particulier au général. L'épreuve comporte 29 items (dont 2 items d'exemple) à résoudre en 8 minutes. Il s'agit de suites numériques dans lesquelles le sujet doit indiquer les deux nombres[1] qui viennent compléter une série proposée.

Exemple d'item :

2-4-6-8-10-12- ?- ?-

➤ Épreuve de Raisonnement analogique (R3)

Cette épreuve présente des situations assez proches des items des matrices de Raven et des tests de facteur *g*. Il s'agit de trouver les lois de transformation

1. Plus précisément il s'agit de sélectionner la bonne réponse parmi 4 réponses possibles.

entre des éléments afin de sélectionner (parmi 5 possibilités) la configuration qui doit compléter la série proposée.

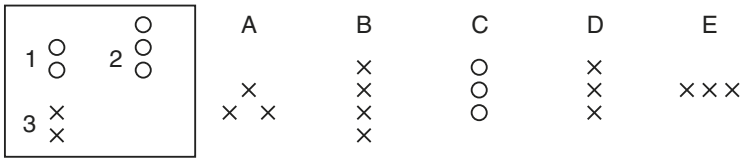

Figure 5.2
Exemple d'item de Raisonnement analogique.

L'épreuve comporte 36 items de ce type (dont 1 item exemple) à résoudre en 10 minutes.

➤ Épreuve de Raisonnement pratique-technique (R4)

Les items reprennent ici des situations pratiques ou techniques : estimation de phénomènes physiques ou mécaniques (suite d'engrenages, par exemple).

Exemple d'item : indiquer l'image qui représente le clou qui s'enfoncera le plus facilement.

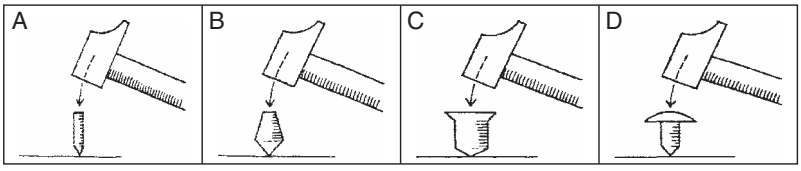

Figure 5.3
Exemple d'item de Raisonnement pratique-technique.

Le sujet doit résoudre 26 items (dont 1 item d'exemple) en 10 minutes. Ces situations sont intéressantes mais le style de graphisme qui date des années 1970 n'est cependant guère attrayant.

➤ Épreuve d'aptitude Spatiale

Elle vise à évaluer les capacités de visualisation spatiale et plus précisément les capacités du sujet à se représenter une configuration en trois dimensions à partir d'un plan en deux dimensions.

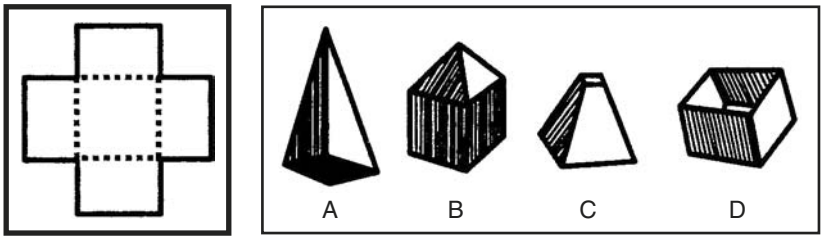

Figure 5.4
Exemple d'item de l'échelle Spatiale.

L'épreuve comporte 42 items (dont 2 items d'exemple) à résoudre en 10 minutes.

➤ Épreuve de Problèmes

Dans laquelle le sujet doit résoudre de courts problèmes arithmétiques, présentés par un énoncé de quelques phrases. Alors que les quatre opérations mathématiques élémentaires sont évaluées plus précisément dans une autre épreuve (*Opérations)* on cherche ici à rendre compte de la capacité du sujet à appliquer des notions mathématiques dans des situations-problèmes. Comme dans l'exemple suivant, le sujet doit sélectionner sa réponse parmi 5 possibilités. Exemple (fictif) d'item :

Exemple

Une corde de 39 m est coupée en trois parties égales. Quelle est la longueur de chaque partie ?

1) A : 14 m
2) B : 12 m
3) C : 13 m
4) D : 23 m
5) E : 10 m

Ce subtest comporte 16 items à résoudre en 6 minutes.

On remarquera ici que lorsque le problème comporte un prix, celui-ci est encore exprimé en francs, et non pas en euros, détail qui renforce l'image obsolète de certaines épreuves.

➤ Épreuve d'Opérations

Vise explicitement à s'assurer de la maîtrise des quatre opérations de base : addition, soustraction, division et multiplication. Pour chaque opération présentée (49 items au total) le sujet doit sélectionner ce qu'il considère comme étant la bonne réponse (5 choix possibles). Temps limité de 10 minutes.

➤ Épreuve d'Attention

Consiste à repérer si un mot, ou groupe de mots, a été correctement recopié. On vise ici à estimer les capacités d'attention et de concentration. Le sujet doit décider si les deux séries sont identiques ou non.

Banque Mondiale Banque Mondial

L'épreuve comporte 55 items (dont 2 items d'exemple) à résoudre en 4 minutes. On peut noter ici la proportion élevée de sélectionner la bonne réponse en se fiant uniquement au hasard (comme dans le subtest R1 : 50 % de chance).

➤ Épreuve d'Orthographe

Comporte 55 items (dont 2 items d'exemple). Le sujet doit indiquer si chaque mot proposé (de langage courant) est correctement orthographié.

Temps limité à 4 minutes. On peut remarquer qu'on ne demande pas au sujet d'écrire l'orthographe correcte du mot mais uniquement de sélectionner la réponse parmi 2 possibles, oui/correcte ou non/incorrecte.

➤ Épreuve de Compréhension verbale

Il s'agit d'indiquer si deux verbes (comme par exemple : ouvrir – fermer) sont semblables ou contraires. Ici encore la probabilité de trouver la bonne réponse par hasard est élevée.

L'épreuve comporte 55 couples de verbes (dont 2 items d'exemple) à résoudre en 4 minutes.

La batterie NV7 comporte donc des épreuves variées, certaines portent sur des aptitudes cognitives de raisonnement (épreuves *R1* à *R4*) et sur des aptitudes spatiales, tandis que d'autres relèvent plutôt des acquisitions scolaires. Chaque épreuve fait l'objet d'un score, la batterie comportant 10 épreuves, le sujet sera donc caractérisé par autant de scores. À ces dix scores

vont se rajouter deux indicateurs composites, EIG (Efficience Intellectuelle Générale) et ES (Efficience Scolaire), ainsi que des indices de rapidité et de précision. Nous présenterons plus loin chacun de ces indicateurs.

Les qualités psychométriques de la batterie NV7

Les études de validation présentées dans le manuel ont été réalisées sur un échantillon de 867 adolescents et jeunes adultes, de niveaux V à VI. Nous en présenterons les éléments principaux.

➤ Analyse de la sensibilité

Les auteurs s'intéressent ici au pouvoir discriminant des épreuves. Les données de 1993 nous indiquent un bon niveau de sensibilité (formes « gaussiennes » des distributions globalement respectées et indicateurs de dispersion satisfaisants) mais il faut être attentif aux points suivants :

- Le subtest *Problèmes* est un peu trop difficile, avec seulement environ 30 % de réussite en moyenne. Ce problème s'est probablement actuellement estompé du fait de l'effet Flynn (*cf.* chapitre 1) ;
- Le subtest *Compréhension verbale* est lui un peu trop facile (58,5 % de réussite), ne permet pas de différencier finement les scores élevés, ce qui est en fait en accord avec les objectifs de cette épreuve qui vise principalement à détecter l'illettrisme (manuel p. 29). Cependant ce problème s'est probablement accentué du fait de l'effet Flynn ;
- Le subtest *Attention* ne présente pas une distribution conforme à une distribution gaussienne, ce qui réduit la sensibilité de ce subtest.

➤ Analyse de la fidélité

Deux méthodes ont été utilisées pour rendre compte de l'homogénéité de chaque épreuve : la méthode pair-impair et l'indice de Kuder-Richardson (KR 20). Le tableau II du manuel (p. 30) nous en donne les valeurs[1].

1. Nous attirons l'attention du lecteur sur le point suivant : ce tableau comporte des erreurs au niveau de l'intitulé des colonnes : la colonne « moyenne » correspond en fait aux indicateurs des corrélations « pair-impair » et la colonne « écart type » correspond aux valeurs des KR20 !

Pour la corrélation pair-impair, les valeurs sont toutes proches de .80, ce qui est acceptable, excepté pour *Raisonnement pratique-technique* avec une valeur plus faible (.57[1]).

Pour les indices KR20, les valeurs s'échelonnent entre .74 et .97, valeurs également acceptables, au moins pour les plus élevées

Recommandation

Les deux épreuves présentant les valeurs de fidélité les plus faibles sont celles de *Raisonnement pratique technique (*corrélation de .74) et de *Problèmes (*corrélation de .75). Les notes obtenues par les sujets dans ces épreuves doivent être considérées comme des évaluations moins précises des aptitudes concernées (manuel, p. 30). Nous avons ici encore un bon exemple de ce que peut apporter à un praticien la lecture attentive du manuel du test : lui fournir les éléments utiles à l'interprétation des résultats et au repérage des points forts et des éventuelles limites de fiabilité de l'épreuve qu'il utilise.

➤ Analyse de la validité

Validité structurale

Pour la validité structurale, on s'attend à observer des corrélations non négligeables entre tous les subtests en raison de l'existence du facteur *g*. Le manuel (tableau IV, p. 31) indique des valeurs de corrélations qui varient de .28 à .70 selon les subtests. Elles vont dans le sens attendu : plus élevées entre les subtests censés évaluer des dimensions voisines et plus faibles lorsqu'il s'agit d'aptitudes plus éloignées. On observe bien, par exemple, une corrélation plus élevée entre deux épreuves de raisonnement (.60 entre *R1* et *R2*) qu'entre une épreuve de raisonnement et une autre épreuve de la batterie (.28 entre *Raisonnement R1* et *Attention*).

En complément de l'analyse simple des corrélations, l'analyse des données est approfondie par les méthodes d'analyse factorielle. En première étape, une analyse est réalisée sans rotation : on retrouve alors un premier facteur général, interprétable comme un facteur *g*, qui explique plus de 56,4 % de la variance. Ensuite, une méthode de rotation Varimax est utilisée, permettant la mise en évidence de trois facteurs, chaque facteur saturant principalement

1. Cette valeur, trop faible selon nous, n'est pas commentée dans le manuel.

une partie des épreuves de la NV7. Le tableau VI du manuel (p. 32) fournit les différentes valeurs de saturation[1].

L'une des applications possibles de ces analyses est de donner des indications concernant ici la construction d'indices composites, c'est-à-dire d'indices combinant différentes épreuves. Les auteurs proposent deux indicateurs composites, EIG et ES, directement reliés à ces facteurs statistiques.

- *Le facteur 1*, qui explique plus d'un tiers de la variance totale, sature particulièrement les épreuves de raisonnement [*R. déductif R1* (.69), *R. analogique R3* (.72) et *R. pratique-technique R4* (.73)] ainsi que l'épreuve *Spatiale* (.80)]. Les auteurs interprètent ce premier facteur comme un facteur de compréhension générale, qui serait relativement indépendant des acquisitions scolaires et du milieu culturel (manuel, p. 32). Ce facteur 1, que l'on pourrait qualifier d'intelligence fluide, fonde la validité du calcul du score composite EIG (Efficience Intellectuelle Générale) ;.
- *Le facteur 2,* avec lui aussi plus d'un tiers de variance, sature plus particulièrement les subtests liés aux acquisitions scolaires : *Opérations* (.84), *Compréhension verbale* (.74), *Orthographe* (.73), et *Problèmes* (.68). Il sature cependant également l'épreuve de raisonnement *Raisonnement inductif R2* (60[2]). Les auteurs rapprochent ce facteur de l'intelligence cristallisée (manuel, p. 32). Il servira de support au calcul du second score composite : le score ES (Efficience Scolaire) ;
- *Le facteur 3* est moins important (il n'explique que 13,8 % de la variance). Il sature principalement le subtest *Attention*, et témoigne de la spécificité de ce qui est évalué dans ce test.

Validité théorique et prédictive

Le manuel ne comporte aucun résultat de recherche concernant la validité prédictive et la validité théorique de la NV7. Nous pouvons supposer que, la NV7 reprenant des tests déjà existants, les auteurs n'ont pas jugé nécessaire de s'assurer de nouveau de leur validité théorique. Néanmoins de telles données mériteraient à notre avis de figurer dans le manuel. Des analyses vis-à-vis de la réussite scolaire seraient également pertinentes à mener.

1. Ici encore une erreur regrettable complique la lecture de ce tableau : la troisième colonne intitulée « facteur 1 » devrait être rattachée à la colonne « après rotation » et non pas à la colonne « avant rotation ».

2. L'épreuve *R2* présente une saturation presque aussi élevée (.58) dans le facteur 1 que dans le facteur 2 (.60). La décision des auteurs de le rattacher exclusivement au score ES mériterait probablement des explications supplémentaires.

Effets du sexe

Enfin, les auteurs fournissent quelques données concernant l'analyse des effets du sexe sur les performances. Sur les dix subtests, on observe cinq différences significatives dans le sens attendu : les garçons obtiennent des scores moyens supérieurs dans les subtests reposant sur du raisonnement concret et/ou sur des aptitudes spatiales (les subtests *Raisonnement technique, Spatial* et *Problème*), les filles obtenant des résultats moyens supérieurs dans l'un des subtests reposant sur des capacités verbales (*Orthographe*) ainsi que dans le subtest *Attention* (d'après les données du tableau III, p. 30 du manuel).

Recommandation

La conséquence de ces différences entre filles et garçons est qu'il conviendra d'utiliser pour ces subtests des étalonnages spécifiques selon le sexe.

La standardisation

➤ La passation

La NV7 est une épreuve de type papier-crayon. Le sujet dispose d'un livret de passation de 64 pages (format A4) dans lequel se trouvent les dix subtests de l'épreuve. Le psychologue dispose d'un manuel qui réunit l'ensemble des consignes.

Rappelons que le psychologue peut décider de ne faire passer qu'une partie seulement des subtests. Chaque subtest se déroule en temps limité (certains subtests sont très courts : 4 minutes, d'autres plus longs : 10 minutes maximum). Pour une passation complète de la batterie il faut compter une durée minimum d'1 heure 45 minutes, exemples compris.

Le sujet dispose d'une feuille de réponse de type auto-scorable sur laquelle il inscrira ses réponses en noircissant les cases correspondantes. Chaque colonne correspond à un subtest.

Recommandation

On peut noter ici que cette feuille n'est pas très attrayante pour le sujet et qu'une erreur de retranscription est possible. Nous ne pouvons que conseiller au praticien d'être très attentif aux éventuelles erreurs à ce

niveau en cherchant, par exemple, à vérifier régulièrement l'exactitude de l'emplacement des réponses données par le sujet aux différents items.

➤ La cotation

La feuille de réponse de type auto-scorable permet une cotation rapide du protocole du sujet (une correction automatisée par lecture optique est également proposée dans le manuel).

Après avoir déplié la feuille de réponse afin de faire apparaître les grilles de cotation, on procède à la correction : on attribue 1 point par réponse correcte, puis on en effectue la somme par colonne afin d'obtenir un score brut pour chaque épreuve.

On reporte ensuite ces dix scores bruts dans la colonne « notes brutes » du tableau d'analyse du profil.

En plus de ces dix scores, le praticien peut calculer les deux indicateurs EIG (Efficience Intellectuelle Générale) et ES (Efficience Scolaire) à partir des formules suivantes (manuel, p. 12) :

Score brut EIG (Efficience Intellectuelle Générale) = *R1 + R3 + R4 + Spatial*

Score brut ES (Efficience Scolaire) = (2 x R2) + (6x Problèmes) + (2x Opérations) + Compréhension verbale + Orthographe

La justification de ces coefficients dans le calcul du score ES n'est pas donnée dans le manuel. Nous pouvons penser qu'ils servent à rééquilibrer le poids de chaque subtest dans le calcul de l'indice ES (car les subtests ne comportent pas le même nombre d'items).

Notre expérience de formation à l'évaluation nous amène à penser que nombre de praticiens effectuent ces calculs avec ces coefficients mais sans réellement en connaître la justification. Pourtant, comme le précisent tant le Code de déontologie que les textes relatifs à l'utilisation de tests[1], le psychologue doit toujours conserver la maîtrise des résultats qu'il est amené à interpréter. Il nous semble alors indispensable que des explications suffisantes soient fournies aux utilisateurs de la NV7 pour qu'ils puissent comprendre le sens et les limites de validité, des calculs qu'ils proposent, et tout particulièrement pour ces deux indicateurs composites ES et EIG.

1. Voir par exemple les *recommandations internationales dans l'utilisation des tests*, qui seront présentées et commentées dans le chapitre 8 de ce livre.

Enfin, le praticien peut également calculer des indicateurs supplémentaires : un indicateur *R* de rapidité et un indicateur *P* de précision, à partir des formules suivantes (manuel, p. 13) :

R = (nombre de réponses produites / nombre de réponses possibles) x 100
P = (nombre de bonnes réponses / nombre de réponses produites) x 100

Ces deux indicateurs R et P fournissent des informations sur les stratégies de réponse du sujet à chaque subtest :

– l'indicateur R, indicateur de rapidité, correspond au pourcentage de réponses données (que ces réponses soient correctes ou non), c'est-à-dire au pourcentage de problèmes abordés par le sujet ;
– l'indicateur *P*, indicateur de précision, correspond au pourcentage d'items correctement résolus parmi ceux ayant été abordés.

Le praticien dispose ici d'indicateurs à la fois quantitatifs et qualitatifs sur les performances du sujet.

Recommandation

Ces scores présentent un réel intérêt pour une approche clinique de l'évaluation réalisée et une restitution à la personne évaluée d'informations sur son propre fonctionnement.

➤ Les étalonnages

Au total, le praticien dispose de douze scores bruts : les dix scores aux subtests et les deux scores composites ES et EIG.

Il dispose également, pour chaque subtest, des scores bruts aux indices R et P.

Chaque note brute doit être transformée en note étalonnée afin de pouvoir être interprétée.

Trois études d'étalonnage sont présentées dans le manuel :

- Une étude de 1990-1991 réalisée dans le cadre de bilans d'orientation sur 300 jeunes faiblement qualifiés (âge moyen 20 ans) fournit 3 étalonnages : pour la population totale et par sexe (manuel, p. 35 à 38) ;
- Une étude de 1991-1992 sur 524 adultes faiblement qualifiés, de niveau maximum BEP (âge moyen 35 ans), dans le cadre d'évaluation en

entreprise (recrutement, promotion) ou de bilans d'orientation, fournit 3 étalonnages : pour la population totale et par sexe (manuel, p. 43- à 49) ;

- Une étude de 1993 sur 104 jeunes apprentis, en cours de préparation d'un diplôme du secteur industriel (CAP, BEP ou bac professionnel). Cet échantillon est fortement masculinisé ce qui explique ici un étalonnage uniquement masculin (manuel, p. 59 à 62) avec distinction possible selon le niveau de diplôme préparé (population totale, préparation CAP/BEP, préparation Bac Professionnel).

Les étalonnages sont cohérents avec les objectifs de la batterie qui, rappelons-le, est destinée à l'évaluation de jeunes adultes et adultes de faible niveau de qualification

Tous les étalonnages sont de type normalisé en neuf classes.

Après avoir déterminé l'étalonnage le plus approprié au sujet évalué, le psychologue va transformer la note brute de chaque subtest en une note étalonnée. Il va ainsi situer le niveau de performance du sujet parmi les neuf catégories proposées. Une évaluation plus globale en cinq niveaux de performance (de -- à ++) est également possible comme indiqué dans le tableau 5.1.

Tableau 5.1

Principes des étalonnages de la NV7.

	Notes étalonnées								
Codage	– –	–		0			+		+ +
Classe	1	2	3	4	5	6	7	8	9
Répartition théorique	4 %	6,6 %	12,1 %	17,5 %	19,6 %	17,5 %	12,1 %	6,6 %	4 %

L'interprétation des scores de la NV7

Les auteurs proposent un guide d'analyse des résultats à la NV7 très utile au praticien pour l'interprétation et la restitution du test.

Ce guide comprend douze étapes allant de la connaissance préalable du sujet à la restitution des résultats (manuel, p. 20 à 24). Les auteurs y précisent également, pour chaque subtest, ce qui est plus spécifiquement évalué. Enfin, ils proposent huit études de cas (p. 24 à 28).

Le praticien dispose ainsi dans le manuel de différentes informations pouvant lui être utiles.

Nous proposons ici une synthèse en quatre points des éléments principaux concernant l'interprétation des différents scores de la NV7.

➤ Analyse de chaque score aux dix subtests

Il s'agit ici de situer le niveau de performance du sujet dans les dix scores étalonnés, par rapport à une population[1] de référence (échantillon d'étalonnage). Le praticien va se référer ici à la feuille de profil afin de repérer les points faibles (score – et --) et les points forts (scores + et ++) du sujet. Dans le cas de codage – on peut parler de niveau très faible, à l'inverse, dans le cas de résultats ++ on peut parler de résultats très élevés.

> **Attention !**
>
> Attention ici à un risque d'erreur : les scores codés 0 ne correspondent pas à des scores bas mais à des scores moyens (par rapport à la population d'étalonnage).

À l'aide des pourcentages de répartition théorique (qui figurent également sur la feuille de profil) le praticien peut situer plus précisément la position du sujet sur les dix scores.

> ***Exemple***
>
> Si le sujet est situé en classe 7, ce score, catégorisé +, peut être considéré comme l'un des points forts du sujet. Plus précisément, ce score étalonné de 7 nous indique que seulement 10,6[2] % de la population de référence[3] dépasse ce niveau de performance, et que 77,3 % des sujets[4] de cette population se situent en dessous de ce niveau.

En complément de cette comparaison interindividuelle (qui précise comment se situent les scores du sujet par rapport aux sujets de l'étalonnage) il est également possible d'adopter une approche intra-individuelle (repérer,

1. Il peut également être judicieux de comparer les résultats d'un même sujet à plusieurs étalonnages afin d'estimer son niveau de performance par rapport à différentes populations de référence.
2. 6,6 + 4 = 10,6 %
3. Il faut bien entendu toujours caractériser cette population de référence (niveau de formation, sexe...).
4. 4 + 6,6 + 12,1 + 17,5 + 19,6 + 17,5 = 77,3 %

par exemple, les propres points forts d'un sujet, c'est-à-dire ses meilleurs résultats parmi les dix subtests).

➤ **Analyse des deux scores composites EIG (Efficience intellectuelle générale) et ES (Efficience scolaire)**

Il s'agit de situer les résultats du sujet sur les deux grandes dimensions synthétiques que sont l'intelligence fluide (représentée par EIG) et l'intelligence cristallisée (représentée par ES). On s'intéresse ici plus précisément :

– au niveau de performance dans chaque indice (en s'inspirant des règles générales que nous venons de présenter) ;
– au décalage éventuel entre EIG et ES. On regardera par exemple dans quelle mesure les capacités du sujet sont dépendantes du contenu, plus ou moins scolaire, des épreuves, ou encore si le niveau des acquis scolaires (ES) reflète bien les potentialités intellectuelles (EIG) ;

Le praticien sera également attentif à l'homogénéité des résultats pris en compte dans le calcul de chacun de ces deux scores : on observera s'il existe un décalage de niveau de réussite dans les subtests constituant chaque indice, ou au contraire, si les résultats sont homogènes (analyse de la dispersion des scores à l'intérieur de chaque indice).

Ces deux indicateurs peuvent également constituer des éléments prédictifs par rapport à un projet de formation : en cas de notes élevées à l'indice ES par exemple, les auteurs conseillent une entrée directe en formation de niveau V, tandis qu'une note faible à cet indice doit inciter le praticien à conseiller plutôt une orientation vers des stages de remise à niveau avant l'entrée éventuelle en formation (manuel, p. 17).

➤ **Analyse des indices de rapidité R et de précision P**

C'est ici l'une des spécificités de cette batterie. Ces deux indices doivent être analysés conjointement afin de fournir des informations sur certaines caractéristiques du sujet. Par exemple un sujet qui présente, sur la majorité des subtests, des scores *R* faibles, mais des scores *P* élevés, est probablement un sujet méticuleux, vérifiant ses réponses, ce qui explique à la fois le faible nombre d'items traités (*R* faibles) mais un pourcentage élevé de bonnes réponses (*P* élevés). On peut ici faire un lien avec les notions de style cognitif, de réflexion/impulsivité (Huteau, 2002). On sera également attentif ici aux éventuelles variations de ces deux indices en fonction des subtests. Ces

aspects de stratégie de réponse pourront être abordés avec le sujet dans la phase de restitution des résultats.

➤ **Analyse de l'adéquation entre les résultats à la NV7 et les projets de formation ou les projets professionnels**

Il ne s'agit pas ici bien entendu de rechercher une stricte adéquation entre profil du sujet et profil du poste et/ou du contenu de la formation, car les résultats de la batterie NV7 (comme plus généralement tout résultat de test) ne sont qu'un des éléments à prendre en compte dans une démarche de conseil (Aubret & Blanchard, 2005). En effet, d'autres facteurs vont intervenir comme l'expérience, la motivation, les intérêts professionnels, la situation familiale... Mais l'analyse de cette adéquation peut être discutée avec le sujet dans la phase de restitution des résultats.

Nous trouvons d'ailleurs dans le manuel des propositions de lecture des résultats en fonction de différents types d'activités professionnelles : par exemple, en analysant conjointement trois subtests [*Raisonnement inductif, Opérations et Problèmes*] le psychologue pourra estimer le degré d'aisance du sujet dans des situations professionnelles nécessitant l'usage de chiffres. Autre exemple, l'analyse conjointe de trois autres subtests [*Raisonnement analogique, Raisonnement pratique-technique et Spatial*] apportera des éléments concernant cette fois les activités professionnelles de type atelier (voir les autres indications du manuel, p. 21 et 23).

Conclusion sur la batterie NV7

Comme les auteurs l'ont souhaité, la batterie NV7 est adaptée à une population de faible niveau de qualification. Ses qualités métriques sont globalement satisfaisantes

Les indicateurs de la performance du sujet sont nombreux, dix scores d'aptitudes et deux scores composites, et permettent une analyse assez complète des aptitudes du sujet.

Le praticien dispose en outre de deux indicateurs, *R* et *P*, qui peuvent apporter des informations utiles pour apprécier le fonctionnement de la personne, informations généralement négligées dans les autres tests.

Le praticien trouvera dans le manuel un bon soutien méthodologique à l'interprétation des résultats, ainsi que des études de cas.

Cela en fait une batterie intéressante pour les niveaux les plus faibles qui mériterait d'être mise à jour pour corriger les quelques erreurs et manques du manuel ainsi que les aspects désuets de certains subtests.

2. La batterie NV5-R

Présentation de la NV5-R

La batterie NV5-R est en partie inspirée de la batterie NV7 mais elle est destinée à des publics de niveau de qualification plus élevé (au minimum équivalent au niveau Baccalauréat). Elle est donc complémentaire, au regard de la population ciblée, de la batterie NV7. Elle est adaptée à un public d'adolescents et d'adultes

Cette batterie, diffusée en 2003 est une version rénovée de la batterie NV5 de 1987. Comme la NV7, la NV5-R est composée d'une combinaison de tests anciens mais les auteurs indiquent que les sous-échelles (ou subtests) ont été sélectionnées en fonction d'une théorie de référence : le modèle du Radex . Cette référence théorique est assez originale et mérite d'être soulignée. Rappelons que, d'après ce modèle, les tests d'intelligence peuvent être positionnés dans un espace bidimensionnel avec en position centrale les tests de facteur *g*. Une présentation synthétique de ce modèle est proposée dans Dickes et Martin (1998) à qui nous empruntons la figure qui illustre ce modèle (voir figure 5.5).

L'interprétation de ce que mesure un test va alors dépendre de sa position sur ce Radex, à partir des principes suivants :

- Plus le test est proche du centre de la figure, mieux il mesure (plus il sature dans) le facteur *g* ; à l'inverse, plus il est situé dans la périphérie et est distant du facteur *g*, et plus il exprimera la mesure d'aptitudes spécifiques ;
- Cette position peut également être interprétée en terme de niveau de complexité : plus un test (une tâche) est intellectuellement complexe, plus il sera situé au centre de la figure ;
- Trois zones peuvent être distinguées dans le Radex, qui correspondent globalement à trois domaines : verbal, spatial et numérique.
- À partir des ces principes d'interprétation, ce modèle en Radex fournit une information sur la liaison du test avec le facteur *g* (niveau de proximité),

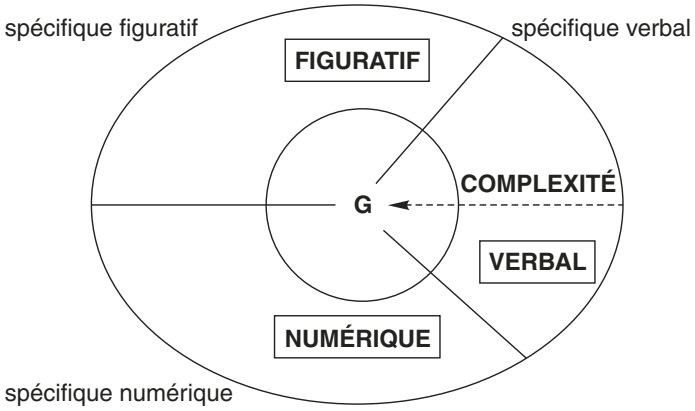

Figure 5.5
Exemple de représentation d'un espace bidimensionnel de type Radex (d'après Dickes et Martin, 1998, p. 31).

ainsi qu'une estimation du domaine évalué plus spécifiquement par l'épreuve (verbal, spatial ou numérique).

Précisons que le manuel comporte une large introduction à ce modèle théorique qui n'est sans doute pas très familier à nombre de psychologues.

Nous verrons plus loin, dans l'interprétation des résultats, qu'un certain niveau de connaissance théorique du modèle de référence est ici particulièrement nécessaire afin de pouvoir réellement maîtriser les indicateurs que l'on peut retirer de cette épreuve. On retrouve, ici encore, la nécessité pour le psychologue de posséder une formation solide, tant au niveau méthodologique qu'au niveau théorique. C'est sans doute ce qui justifie la partie importante consacrée dans le manuel de la NV5-R à la présentation théorique du modèle en Radex.

La batterie NV5-R regroupe neuf épreuves :

1. Raisonnement général,
2. Raisonnement inductif,
3. Raisonnement spatial,
4. Raisonnement pratique/technique,
5. Compréhension verbale,
6. Vocabulaire,
7. Orthographe,

8. Calcul,
9. Attention.

Comme pour la NV7, on peut remarquer ici que certaines épreuves évaluent un raisonnement et/ou des aptitudes, alors que d'autres épreuves relèvent plutôt de connaissances scolaires (comme par exemple *Orthographe* ou *Calcul*).

Pour chaque épreuve, ou subtest, on dispose de consignes et d'étalonnages séparés, ce qui offre une souplesse d'utilisation (le praticien, par exemple, peut ne faire passer qu'une partie des épreuves).

Enfin, comme nous allons le voir, certaines de ces épreuves sont directement issues de la NV7 avec parfois des modifications concernant les temps de passation (afin sans doute de rendre les épreuves plus difficiles[1].)

Détaillons maintenant chacune de ces 9 épreuves :

➤ Épreuve de Raisonnement général

Comporte 49 items (dont 7 exemples) diversifiés tant au niveau de la tâche (on y trouve plusieurs types de raisonnement), qu'au niveau du support (numérique, verbal...), ceci afin de proposer dans une même épreuve un large éventail de situation. L'objectif ici étant bien d'évaluer un raisonnement général, proche de la notion de facteur *g*, mais également d'évaluer la capacité du sujet à faire preuve de flexibilité cognitive (ou dynamisme intellectuel), définie comme la capacité du sujet à s'adapter à des changements dans le type de tâche proposé.

Les exemples suivants donnent un aperçu de la diversité des items de ce subtest :

1. Des items de type « chercher l'intrus », exemple :
Désignez parmi les 5 mots suivants celui qui ne fait pas partie de la série :
Bois – Bouchon – Pierre – Bateau – Liège
2. Des items de type loi de série sur support numérique, dans lesquels le sujet doit poursuivre une suite proposée ; exemple fictif :
2-4-6-8-10-??-??
3. Des items de logique verbale ; exemple fictif :
Julie est plus petite que Fabienne, Sylvie est plus petite que Julie, par conséquent Fabienne est la plus grande des 3 ?
4. Des items dans lesquels le sujet doit montrer sa compréhension de dictons.

1. Le manuel de la NV5 R ne donne pas de précisions à ce sujet. Il nous semble pourtant important que l'origine des items et/ou des subtests soient précisée.

Ce subtest est assez proche du test BV9 de Bonnardel[1].

Le sujet dispose de 20 minutes pour réaliser l'épreuve. C'est d'ailleurs l'épreuve la plus longue de la batterie. Le nombre assez conséquent d'items, et la durée de passation, font de ce subtest une réelle épreuve indépendante.

Concernant les modalités de réponse, la même limite apparaît que celle évoquée à propos de certains subtests de la NV7 : pour certains items les possibilités de réponse (de type QCM) ne sont pas assez nombreuses et la probabilité de trouver la bonne réponse au hasard est trop élevée.

➤ Épreuve de Raisonnement spatial

Le sujet doit se représenter une configuration en trois dimensions à partir d'un plan en deux dimensions. Il s'agit en fait de la même épreuve que celle qui est présente dans la NV7[2] avec 40 items à résoudre en 8 minutes (au lieu de 10 pour la NV7).

➤ Épreuve de Compréhension verbale

Elle comporte 12 items à résoudre en 8 minutes. Le sujet doit indiquer quelles sont les deux phrases (parmi quatre possibilités) qui sont le plus proches d'une pensée (exprimée sous forme de dicton). Exemple d'items :

> **On a besoin d'un plus petit que soi.**
> 1. Il faut regretter que ce soient toujours les plus forts qui l'emportent sur les faibles.
> 2. Ne négligeons pas l'aide que peuvent nous apporter les faibles.
> 3. L'appui des humbles est parfois utile aux grands.
> 4. Petit enfant deviendra grand.

➤ Épreuve de Calcul

Cette épreuve est directement issue de l'épreuve *Opération* de la NV7 avec ici 48 items à résoudre en 10 minutes.

---

1. Certains items semblent d'ailleurs largement inspirés du BV9 de Bonnardel, sans que les auteurs le précisent explicitement.
2. Comme pour tous les subtests issus de la NV7 nous renvoyons le lecteur aux exemples d'items donnés dans la partie précédente (NV7).

➤ Épreuve de Raisonnement pratique/technique

Elle ne comporte pas les mêmes items que le subtest de la NV7 mais en est assez proche, aussi bien au niveau du type de support que, malheureusement, au niveau du type de graphisme. Le sujet doit résoudre ici 33 items en 9 minutes.

➤ Épreuve de Raisonnement inductif

Destinée à évaluer la capacité du sujet à raisonner du particulier au général, elle comporte 27 items, dont 3 exemples, à résoudre en 8 minutes. Il s'agit ici encore d'une épreuve (R2) de la NV7 dans laquelle le sujet doit découvrir les lois de progression de séries numériques.

➤ Épreuve d'Attention

Elle aussi est issue de la NV7 et comporte 52 items, avec un temps de passation de 3 minutes.

➤ Épreuve de Vocabulaire

Comporte 56 items, à résoudre en 4 minutes. La tâche consiste à sélectionner parmi 3 mots proposés les deux mots qui sont soit de même sens, soit de sens opposé. On cherche à évaluer la connaissance du vocabulaire. Exemple d'items :

> **1. Grand / 2. Sec / 3. Vaste**
> **Réponses possibles : 1 et 2 ; 1 et 3 ; 2 et 3.**

➤ Épreuve d'Orthographe

Le sujet doit indiquer si le mot présenté est correctement orthographié. Cette épreuve comporte 54 items, à résoudre en 3 minutes. Il n'est pas demandé au sujet d'orthographier correctement le mot mais d'indiquer si le mot présente, ou non, une erreur. Ici encore la probabilité de trouver la bonne réponse au hasard n'est pas négligeable. Exemple d'items :

> **Le tiroire**

Les qualités psychométriques

Le manuel détaille les procédures de révision des épreuves de la NV5 qui ont conduit à la NV5-R (révisée) : révision de la notation, analyse des biais d'items...

La phase d'expérimentation de la NV5-R a été effectuée auprès d'un échantillon de 460 sujets, âgés de 17 à 57 (moyenne de 26 ans), de niveau d'étude du CAP à Bac + 2.

➤ Analyse de la sensibilité

Les taux moyens de réussite des subtests varient entre 33 % (pour le raisonnement pratique-technique) à 77 % (pour le raisonnement inductif). On observe donc une variabilité assez importante des subtests de la batterie NV5-R : ils ne sont pas tous de même niveau de difficulté. L'analyse des dispersions montre que les scores bruts ne se répartissent pas tous selon une courbe gaussienne. C'est sans doute ce qui explique l'utilisation d'étalonnages par déciles (voir plus loin).

Comme attendu, les taux de réussite varient en fonction du niveau d'études.

➤ Analyse de la fidélité

La fidélité est évaluée à partir de l'indice d'homogénéité interne *alpha* de Cronbach et de l'indice KR 20. Les valeurs prises pour ces indices pour chaque épreuve figurent dans le tableau 5.2.

Tableau 5.2

Indices de fidélité interne de la NV5-R (d'après le manuel, p. 41-43).

Subtests	R. général	R. spatial	Comp. verbale	Calcul	R. pratique	R. inductif	Attention	Voc.	Orth.
Alphas de Cronbach	0,86	0,88	0,83	0,85	0,75	0,89	0,94	0,94	0,88

Les alphas variant de .75 à .94, nous pouvons considérer l'homogénéité interne de la NV5-R comme satisfaisante. La valeur relativement modérée observée pour l'épreuve de raisonnement pratique-technique (.75) doit nous inciter néanmoins à la prudence dans l'interprétation des résultats à ce subtest (manuel, p. 40).

L'erreur standard de mesure est variable selon les subtests, mais il faut signaler ici que le manuel fournit une estimation de cette erreur pour chaque score possible dans certains subtests (voir tableau 12, p. 45 du manuel).

➤ Analyse de la validité

Une analyse statistique de la validité structurelle de la NV5-R, permet de situer les subtests sur une structure en Radex. Rappelons que ce modèle de référence (le radex) est assez peu utilisé dans les tests, et sans doute peu familier à nombre de psychologues, mais que le manuel est bien documenté à ce sujet.

Les auteurs cherchent alors à savoir si leurs données sont bien conformes à ce modèle théorique. Par un traitement statistique particulier (analyse par échelonnement multidimensionnel) on peut observer que la configuration des subtests de la NV5-R est bien compatible avec le modèle théorique supposé. On observe, par exemple, un positionnement central de l'épreuve de *Raisonnement général*, et un positionnement dans la zone attendue pour les épreuves reposant plutôt sur un contenu verbal.

Les subtests se répartissent également en fonction de leur niveau de généralité. Comme attendu, les tâches les plus complexes apparaissent vers le centre et les tâches les plus spécifiques en périphérie. Les détails de la structure observée figurent dans le manuel (voir en particulier la figure 10, p. 39). Ces données apportent des éléments de validité interne de l'épreuve. Par contre, aucun élément d'information ne nous est donné sur le niveau des liaisons entre les différents subtests de la NV5- R.

Comme pour la NV7, aucun résultat d'expérimentations concernant la validité prédictive de l'épreuve ou encore la validité de chaque subtest avec une autre épreuve (validité concourante) n'est malheureusement présenté dans le manuel. Il serait par exemple pourtant utile de disposer de données d'études comparant les résultats du subtest *Raisonnement général* avec ceux d'une épreuve de type facteur *g*.

Recommandation

Nous signalons cependant un article postérieur au manuel (Thiébaut *et al.*, 2005) qui apporte des éléments d'information sur les qualités prédictrices de cette batterie NV5-R. L'article présente les résultats d'une recherche, menée à la demande d'une entreprise, visant à analyser l'efficacité de

ses méthodes de recrutement. Utilisée dans un dispositif de sélection de vendeurs amenés ensuite à suivre une formation, la NV5-R, et plus particulièrement les subtests *Raisonnement général, Compréhension verbale* et *Orthographe*, s'avèrent être de bons prédicteurs du niveau de réussite dans cette formation. Nous renvoyons le lecteur intéressé par ces aspects à la lecture de cet article.

La standardisation

➤ La passation

La NV5-R est une épreuve de type papier-crayon, utilisable en individuel ou en collectif.

Le matériel se compose d'un manuel pour le psychologue (de 108 pages), de cahiers de passation et de feuilles de réponse auto-scorables. Chaque subtest se déroule en temps limité (de 3 à 20 minutes selon les subtests) avec au total un temps d'environ 2 heures si le sujet passe toutes les épreuves.

La feuille de réponse est organisée comme celle de la NV7 : le sujet doit inscrire ses réponses en noircissant les cases correspondantes sur une feuille de réponse de type auto-scorable. Chaque colonne correspond à un subtest.

Recommandation

Nous signalons, comme pour la NV7, que cette feuille n'est pas très attrayante pour le sujet et qu'une erreur de retranscription est possible. Nous ne pouvons que conseiller au praticien d'être très attentif à l'exactitude de l'emplacement des réponses du sujet.

➤ La cotation

Après avoir déplié la feuille de réponse afin de faire apparaître les grilles de correction, on procède à la cotation. On accorde 1 point par bonne réponse (sauf cas particuliers[1] signalés dans le manuel). Le psychologue additionne les points obtenus dans chaque subtest (chaque colonne) afin d'établir les

1. Il s'agit d'items dans lesquels le sujet doit donner 2 réponses : on accordera alors 1 point si – et seulement si – les 2 réponses sont correctes.

neuf scores bruts. Il reportera ensuite ces neuf scores bruts dans le tableau « Profil détaillé » qui figure en haut de la feuille de profil.

➤ Les étalonnages

Ils permettent de transformer les scores bruts en notes étalonnées. Le manuel propose ici plusieurs étalonnages :

– un étalonnage *hétérogène*, sur l'échantillon total de 632 sujets, avec trois possibilités : total, garçons, filles ;
– des étalonnages *par niveau d'études* : niveau d'études inférieur au baccalauréat (139 sujets), égal au bac (258 sujets) et supérieur au Bac (212 sujets). Par contre on ne dispose pas ici de données séparées selon le sexe.

Tous ces étalonnages sont de type décilage[1]. Ce choix est expliqué par le fait que les distributions des scores ne respectent pas suffisamment la courbe de Gauss pour établir des étalonnages standardisés.

Recommandation

L'utilisateur prendra soins de ne pas confondre l'interprétation d'un étalonnage par décilage avec l'interprétation d'un étalonnage normalisée, comme celui, par exemple, de la NV7.

Une fois l'étalonnage sélectionné (étalonnage hétérogène ou par niveau d'études) le psychologue doit donc transformer les neufs scores bruts et définir les neuf notes étalonnées qu'il reportera sur la feuille de profil (scores variant de 1 à 10). Comme pour la NV7, cette feuille lui permettra de repérer rapidement les forces et faiblesses du sujet.

À partir des notes étalonnées, il peut également calculer des scores composites.

1. Rappelons que dans un étalonnage de ce type chaque classe représente 10 % de l'effectif.

Recommandation

Attention pour le calcul des scores composites : il s'agit bien ici d'utiliser les notes étalonnées et non pas, comme dans le cas de la NV7, les scores bruts.

Deux types de scores composites sont ici envisagés :

– les notes du profil d'aptitude,
– les notes du profil cognitif.

- Pour déterminer le *profil d'aptitudes*, on regroupe les épreuves relevant des mêmes dimensions afin d'obtenir un score en aptitude verbale, en aptitude spatiale et en aptitude numérique, selon les indications du manuel. À ces trois indices va se rajouter l'aptitude générale (épreuve de *Raisonnement général) ;*
- Pour déterminer le *profil cognitif*, on se réfère au modèle théorique de référence, le modèle en Radex, afin de déterminer trois scores :
  – l'un relevant des capacités du sujet face à des tâches générales (et complexes),
  – le second reposant sur des tâches de niveau intermédiaire,
  – le dernier relevant de tâches spécifiques.

Nous reprenons ci-dessous le détail de chaque profil en indiquant les subtests de rattachement :

1. Profil d'aptitude

– Aptitude générale : Raisonnement général.
– Aptitude verbale : Compréhension verbale + vocabulaire + attention + orthographe (et diviser cette somme par 4).
– Aptitude spatiale : Raisonnement spatial + Raisonnement pratique technique (et diviser cette somme par 2).
– Aptitude numérique : *Raisonnement inductif* + *calcul* (et diviser cette somme par 2).

2. Profil cognitif
- Général : Raisonnement général.
- Intermédiaire : Compréhension verbale + vocabulaire + Raisonnement spatial + Raisonnement inductif (et diviser cette somme par 4).
- Spécifique : attention + orthographe + Raisonnement pratique technique + calcul (et diviser cette somme par 4).

Rappelons que chaque note de profil est établie à partir des notes étalonnées des subtests. Par la division du total de ces notes on obtient alors directement une note de profil étalonnée, comme les subtests, de 1 à 10.

Au final, le praticien peut disposer des indicateurs étalonnés suivants :

- 9 notes de subtests ;
- 4 notes du profil d'aptitudes : aptitude générale, aptitude verbale, aptitude spatiale et aptitude numérique ;
- 3 notes du profil cognitif : général, intermédiaire et spécifique.

Les bases d'interprétation des scores

Comme pour la NV7, le manuel de la NV5-R propose un support très appréciable à l'interprétation des résultats et des profils observés. Il comprend huit pages dédiées à l'interprétation des différents scores (p. 71 à 78), ainsi qu'une dizaine de pages consacrés à la présentation de quatre études de cas (p. 79 à 90).

L'interprétation proposée se fait dans un premier temps au niveau des subtests, puis dans un second temps au niveau des scores composites. La particularité de l'épreuve est qu'elle peut fournir deux types de scores composites (différentes combinaisons d'items) en lien direct avec les deux cadres de référence théorique proposés (analyse classique en aptitudes ou référence au modèle du Radex).

Quel que soit le niveau d'analyse (subtest ou scores composites), rappelons ici que nous disposons d'étalonnages de type décilages, qui comportent 10 % de sujets dans chaque groupe. On considérera un score égal ou inférieur à 3 comme un score faible, et un score égal ou supérieur à 8 comme un score élevé. Les scores compris entre 4 et 7 inclus étant considérés comme des scores moyens (obtenus par 40 % de la population de référence). Le

praticien trouvera dans le manuel des informations sur ce qui est évalué plus précisément dans chaque subtest (p. 71 à 75).

L'analyse des neuf subtests permet de dresser le profil du sujet, de cerner ses points forts et ses points faibles. On procédera, comme pour la NV7, à une analyse interindividuelle (comment se situent les neufs scores d'un sujet par rapport l'étalonnage ?) mais également intra-individuelle (repérer, par exemple, ses points forts, c'est-à-dire ses meilleurs résultats parmi les neuf subtests).

L'analyse du profil d'aptitudes prend la forme, assez classique, d'une interprétation des résultats du sujet en fonction de trois domaines : aptitude verbale, aptitude spatiale et aptitude numérique. La note d'aptitude générale, composée, rappelons-le, uniquement du subtest *Raisonnement général*, peut être considérée comme l'expression du niveau de facteur *g*.

Le psychologue pourra à cette étape faire des liens entre le profil d'aptitudes du sujet et ses projets de formation et/ou ses projets professionnels, dans la même logique d'interprétation des résultats que celle exposée dans la présentation de la batterie NV7.

Pour **l'analyse du profil cognitif**, l'interprétation des scores doit se faire en relation avec la théorie de référence : le modèle en Radex.

La note *générale,* qui correspond en fait à la note d'aptitude générale du profil d'aptitudes, est ici interprétée comme l'indicateur des capacités du sujet à résoudre des tâches générales (que l'on peut retrouver dans un grand nombre de situations) et complexes.

La note *intermédiaire* va rendre compte des capacités du sujet face à des tâches un peu moins complexes. Enfin, la note *spécifique* est relative aux tâches spécifiques, relativement simples, le plus souvent réduites à l'application de règles.

Les quatre études de cas présentées dans le manuel permettent d'illustrer les grandes lignes d'interprétation des résultats selon les différents niveaux d'analyse (subtests ; profil aptitudes ; profil cognitif).

Le psychologue pourra éventuellement analyser les résultats du sujet en référence aux profils de réponse caractéristiques de quatre groupes de sujets identifiés dans le manuel (voir p. 48 à 56). En ce qui nous concerne, nous ne trouvons qu'un intérêt relatif à cette possibilité de comparaison.

Conclusion sur la NV5 R

La batterie NV5-R est une batterie assez difficile, adaptée à des sujets de niveau minimum Baccalauréat.

Elle regroupe des subtests assez variés et comprend une mesure fiable de l'intelligence générale (ou facteur *g*) par le subtest *Raisonnement général*. Le psychologue peut d'ailleurs, s'il le souhaite, n'utiliser dans un premier temps, que ce subtest, afin d'établir une estimation du niveau général du sujet. Puis, par la suite et en fonction des besoins, utiliser les autres subtests de l'épreuve.

Cette batterie permet d'obtenir neuf scores, représentatifs de neuf aptitudes distinctes, ainsi que des indices composites, certains assez classiques (les quatre scores du profil d'aptitudes), d'autres plus originaux (les trois scores du profil cognitif).

L'une des spécificités de la NV5-R est qu'elle repose sur un modèle théorique assez peu utilisé en psychométrie : le modèle en Radex. Ce modèle, largement développé dans le manuel, permet de caractériser le sujet par son profil *cognitif.* Cette possibilité d'interprétation des scores vient s'ajouter à l'interprétation classique en terme d'aptitudes.

Le manuel gagnerait à être complété d'études concernant la validité prédictive de la batterie (mais comme nous l'avons indiqué le lecteur pourra consulter l'article de Thiébaut *et al.*, 2005). Il fournit, par contre, des indications pertinentes ainsi que des études de cas apportant une aide utile dans l'interprétation des résultats.

3. La batterie DAT 5

Présentation

La batterie DAT 5 (*Differential Aptitude Tests* : Tests Différentiels d'Aptitudes) est issue de l'épreuve DAT qui a été publiée la première fois en 1947 aux États-Unis. Cette 5e édition DAT5 est la version française de la dernière révision de l'épreuve, éditée aux États-Unis en 1990, et adaptée en France par les ECPA dans les années 1998-2000, avec une diffusion en 2002. C'est une épreuve très utilisée aux États-Unis.

Avant la publication de cette dernière version nous ne disposions en France que de la première version de 1974.

La batterie DAT 5 se différencie des batteries NV7 et NV5-R que nous venons de présenter, par la variété des publics auxquels elle convient : elle est utilisable aussi bien pour des publics scolaires (quatre étalonnages scolaires sont disponibles : niveau 3e, enseignement professionnel, 2e, 1ère et Terminale) que pour des publics adultes (deux étalonnages : niveau CAP/BEP et niveau Bac).

Comme pour la première version, la DAT 5 se compose de huit épreuves :

- Trois sont présentées comme mesurant les aspects principaux de l'intelligence (*Raisonnement Verbal*, *Raisonnement Numérique*, *Raisonnement Abstrait*) ;
- Deux subtests évaluent des aptitudes plus spécifiques (*Raisonnement Mécanique* et *Relations Spatiales*) ;
- Deux subtests évaluent plutôt des connaissances (*Orthographe* et *Grammaire*) ;
- Une épreuve de rapidité et de précision (*Vitesse de Perception et Précision*).

On remarquera qu'il s'agit ici, comme dans les batteries NV5-R et NV7, d'évaluer des capacités cognitives (*Raisonnement)* mais également des connaissances scolaires (exemple : *Orthographe...*).

Comme pour les autres batteries factorielles, le psychologue dispose de consignes et d'étalonnages séparés pour chaque subtest, ce qui autorise une grande souplesse d'utilisation.

La passation complète de la batterie nécessite un temps d'environ 2 heures 10 minutes à 2 heures 30 minutes.

Reprenons chacune des huit épreuves.

➤ L'épreuve de Raisonnement Verbal (R.V.)

Elle comporte 25 items dans lesquels le sujet doit compléter des analogies. Le sujet doit choisir les deux termes qui conviennent le mieux pour compléter l'analogie[1] présentée (parmi cinq possibilités). Le temps de passation est limité à 18 minutes.

L'exemple suivant permet de bien comprendre la tâche demandée :

> **... est à *aboyer* ce que *chat* est à ...**
> **Réponse A : miauler...chaton**

1. Sternberg a proposé une analyse (composantielle) de ce type de tâche (Huteau & Lautrey, 1999, p. 214).

Réponse B : chien...miauler
Réponse C : chien...griffer
Réponse D : chien...chaton
Réponse E : se réfugier...griffer

➤ L'épreuve de Raisonnement Numérique (R.N.)

Cette épreuve comprend 25 items à résoudre en 20 minutes. Ces items reflètent divers types de problèmes (suites numériques, estimation de calculs, équations...). Il s'agit ici d'évaluer la capacité à raisonner à partir de données numériques et non pas seulement la capacité à effectuer des calculs. Le sujet dispose de 20 minutes. Exemple d'item :

Quel chiffre peut remplacer le ? dans cette addition ?

5 ?
+ 2
= 58

Réponse A : 3
Réponse B : 4
Réponse C : 7
Réponse D : 9
Réponse E : Aucun

➤ L'épreuve de Raisonnement Abstrait (R.A.)

Elle porte sur les capacités de raisonnement non verbal. Les 25 items reprennent des taches typiques de tests de type facteur *g* : chercher les règles de transformation et les appliquer afin de trouver une figure qui vient compléter une série proposée. Le temps est limité ici à 15 minutes.

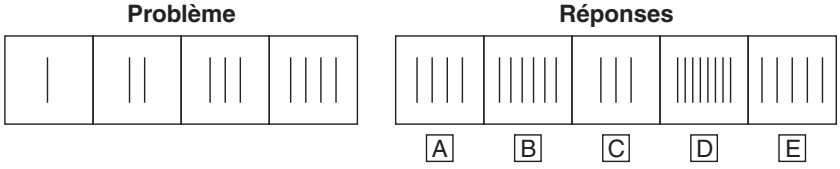

Figure 5.6
Exemple d'item de Raisonnement Abstrait.

Ces trois épreuves de raisonnement (*Verbal, Numérique et Abstrait*), sont présentées comme évaluant les principaux aspects de l'intelligence générale

(manuel DAT 5, p. 5). Tout en étant des épreuves indépendantes, elles sont regroupées dans un même cahier de passation. Elles constituent le noyau central de la DAT 5 et permettent d'évaluer les capacités de raisonnement à partir de trois types de support : verbal, numérique et non verbal.

Les autres épreuves de la DAT 5 évaluent des aspects de l'intelligence considérés comme plus spécifiques (le raisonnement mécanique, l'aptitude spatiale et la vitesse de perception) ou des aspects plus liés aux connaissances en Français (Orthographe et Grammaire).

Les huit épreuves de la DAT 5 ne se situent donc pas sur le même plan par rapport à l'importance des aptitudes évaluées.

Les autres épreuves de la DAT 5

➤ L'épreuve de Raisonnement Mécanique (R.M.)

Le sujet doit résoudre des situations assez concrètes comparables aux items présentés dans le subtest *Raisonnement Pratique-technique* de la NV7 (problèmes de phénomènes physiques, d'engrenages, de poulies...). L'épreuve comporte 30 items à résoudre en 15 minutes. Les items de type QCM ne comportent que trois possibilités de réponse ce qui semble insuffisant (probabilité non négligeable de sélectionner la bonne réponse au hasard).

➤ L'épreuve de Relations Spatiales (R.S.)

Elle comporte 30 items dans lesquels le sujet doit sélectionner la figure qui serait obtenue si une figure modèle était pliée. Cette épreuve porte principalement sur les capacités de représentations mentales de figures géométriques. Le temps est ici limité à 15 minutes.

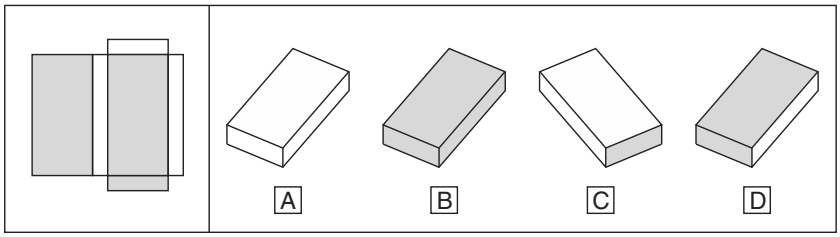

Figure 5.7
***Exemple d'item du subtest* Relations Spatiales.**

➤ L'épreuve de Vitesse de Perception et Précision (V.P.P.)

Ici, le sujet doit résoudre une tâche perceptive simple, le plus vite possible. Il s'agit de retrouver sur la feuille de réponse la combinaison de deux lettres (ou chiffres) qui est soulignée sur le cahier.

Cette épreuve comporte deux séries de 100 (temps de 3 minutes pour chaque série), mais seule la deuxième série interviendra dans la notation. Exemple d'item :

> AB AC <u>AD</u> AE AF
> propositions de réponses : AC AE AF AB AD

➤ L'épreuve d'Orthographe

Cette épreuve consiste à repérer le mot qui est écrit de façon incorrecte parmi quatre mots présentés. Il s'agit d'évaluer les connaissances orthographiques à partir de mots français assez courants. Cette épreuve comporte 30 items à résoudre en 8 minutes. Exemple d'item :

> **A : papier**
> **B : soleille**
> **C : chaise**
> **D : agréable**

➤ L'épreuve de Grammaire

Ici le sujet doit indiquer dans quelle partie d'une phrase se trouvent éventuellement des fautes de grammaire, de conjugaison ou de majuscule. Il ne s'agit pas de corriger ces fautes mais uniquement d'indiquer l'endroit où elle est située (ou indiquer l'absence d'erreur). L'épreuve comporte 30 items à résoudre en 12 minutes. Exemple d'item :

> **Nous serat-il / possible d'aller / travailler la / semaine prochaine.**
> **A B C D**

La DAT 5 est donc composée de huit épreuves, et permet de caractériser le sujet selon les huit dimensions correspondant à ces épreuves. Le calcul d'un indicateur d'efficience scolaire (score composite) est également possible en combinant les résultats obtenus à l'épreuve de Raisonnement Verbal et Raisonnement Numérique (voir plus loin).

Les qualités psychométriques de la DAT 5

Rappelons que la version française DAT5 est une version adaptée de la DAT 5 américaine éditée en 1990 aux États-Unis. Le manuel contient une description des conditions d'élaboration de cette version américaine qui comprend deux niveaux :

– niveau 1 (correspondant à des élèves de 5e à la 3e) ;
– niveau 2 (correspondant à des élèves de la 2e au Bac).

Pour élaborer la version française, la procédure classique d'adaptation d'une épreuve a été suivie : traduction ou création de nouveaux items, expérimentation et développement des formes définitives.

Il n'est pas inutile de donner les grandes lignes de cette adaptation.

Sur cette version américaine le manuel ne fournit que très peu de données. On apprend juste que les coefficients de fidélité KR-20 varient de .82 à .95, ce qui démontre un bon niveau de consistance interne, que des corrélations très élevées (entre .86 et .90) ont été observées avec des tests d'aptitude (sans nous préciser de quels tests il s'agit) et que les subtests de la DAT5 peuvent être considérés comme de « bons prédicteurs de la réussite scolaire » (manuel DAT5, p. 26), sans nous donner davantage de précision.

La version française est directement adaptée de cette version américaine, mais avec deux particularités :

– d'une part, seul le niveau 2 a été adapté ;
– d'autre part, elle en constitue une forme abrégée (réduction du nombre d'items afin de diminuer le temps de passation).

Les auteurs indiquent que l'objectif central de cette adaptation a été de rester au plus près de la version originale. Dans la plupart des cas, les items originaux ont donc été traduits. Lorsqu'une simple traduction n'était pas pertinente (par exemple pour l'épreuve d'orthographe), de nouveaux items ont été créés. Au final la forme française expérimentale comportait de 45 à 67 items selon les subtests.

Cette épreuve a été testée auprès d'un public scolaire (2 651 élèves de niveau 3e à terminale, avec une partie de l'échantillon scolarisé dans l'enseignement professionnel) ainsi que sur un échantillon d'adultes (212 adultes, de niveau CAP à Baccalauréat).

Pour aboutir à une version réduite, la sélection des items de l'épreuve définitive a été réalisée à partir de quatre critères principaux : le pouvoir

discriminant des items, la pertinence des distracteurs, la typologie des items et le niveau global de difficulté.

Le tableau 5.3 résume les étapes de cette adaptation.

Tableau 5.3
Nombre d'items des différentes formes de la DAT 5.

Subtests de la DAT 5	Forme définitive américaine	Forme expérimentale française	Forme définitive française
Raisonnement Verbal	40	50	25
Raisonnement Numérique	40	60	25
Raisonnement Abstrait	40	45	25
Relations Spatiales	50	55	30
Raisonnement Mécanique	60	67	30
Orthographe	40	60	30
Grammaire	40	60	30

Comme nous l'indique le tableau, si la version finale française est bien une version abrégée elle comporte un nombre suffisant d'items (de 25 à 30) dans chaque subtest pour garantir un certain niveau de fiabilité.

➤ Analyse de la sensibilité

Sur l'échantillon scolaire

Les valeurs moyennes de réussite, ainsi que les valeurs de dispersion, indiquent que globalement l'épreuve est bien adaptée à ce type de public. Dans l'ensemble, on observe bien une évolution des scores moyens en fonction des niveaux scolaires. Cette augmentation n'a cependant pas été observée pour les élèves de Terminales, ce qui explique que les niveaux 1re et Terminales ont été regroupés au sein du même étalonnage.

Sur l'échantillon adulte

Les caractéristiques moyennes de réussite sont également satisfaisantes avec, comme attendues, des différences significatives selon les niveaux de qualification (CAP ou Baccalauréat).

➤ Analyse de la fidélité

Sur l'échantillon scolaire

Les indicateurs *alpha* de Cronbach varient de .74 à .96 selon les subtests, ce qui témoigne d'un niveau global satisfaisant d'homogénéité interne.

La fidélité test-retest a été estimée à partir d'un échantillon d'élèves de 3e. Les coefficients de corrélation varient entre .56 et .86 selon les subtests. Certaines valeurs sont donc un peu faibles (*Vitesse de précision* .56 et *Raisonnement Abstrait* .58).

Signalons que les valeurs caractéristiques (moyennes, écart-type), les coefficients *alpha* de Cronbach, ainsi que les erreurs de mesure figurent, pour chaque niveau scolaire, en annexe du manuel.

Sur l'échantillon adulte

Les coefficients *alpha* varient ici de manière similaire à l'échantillon scolaire (de .77 à .98 selon les subtests) mais la fidélité test-retest n'a pas été évaluée. Les erreurs de mesure sont également indiquées pour chaque niveau de qualification.

➤ Analyse de la validité

Sur l'échantillon scolaire

Une première analyse de validité porte sur la structure de l'épreuve. Les coefficients de corrélation entre les huit subtests varient de .06 (entre *Orthographe* et *Raisonnement Mécanique*) à .65 (entre *Grammaire* et *Orthographe*). On retrouve globalement le *pattern* attendu : corrélations les plus élevées entre des tests évaluant des dimensions les plus proches (des valeurs autour de .65 par exemple entre les trois principales épreuves de raisonnement : *Raisonnement verbal, numérique et abstrait*) et corrélations plus faibles entre des tests évaluant des aptitudes plus éloignées. Le psychologue trouvera dans le manuel toutes les valeurs des intercorrélations (tableau 7.6, p. 58).

Une deuxième étude de validité porte sur la liaison entre la forme DAT de 1974 et la forme DAT 5. L'échantillon est malheureusement assez restreint (une cinquantaine d'élèves de 3e) ce qui peut contribuer à expliquer la faiblesse de certaines des valeurs observées. En effet, les corrélations varient entre .77 et .43 comme nous l'indique le tableau 5.4.

L'obtention de valeurs inférieures à .70 peut paraître surprenante car il s'agit bien ici de comparer deux versions différentes (1974 et 2002) du même test. Cet aspect est signalé dans le manuel :

Tableau 5.4
Corrélations entre les subtests de la DAT et les subtests de la DAT 5.

Raisonnement Verbal	.77
Raisonnement Numérique	.43
Raisonnement Abstrait	.45
Relations Spatiales	.55
Raisonnement Mécanique	.71
Orthographe	.65
Grammaire	.50

« Certaines corrélations obtenues ne sont pas aussi élevées que ce que l'on pouvait attendre. » (manuel DAT 5, p. 59).

Les auteurs du manuel proposent une explication qui repose sur les évolutions temporelles entre les deux versions au niveau du contenu des items et des échantillons. Sans vouloir négliger ces explications, la faiblesse de certaines corrélations (par exemple,.45 pour le Raisonnement Abstrait) devrait amener les auteurs à envisager une autre expérimentation sur un échantillon plus vaste de sujets.

Enfin une dernière étude de validité, validité critériée, porte sur les liaisons entre DAT5 et résultats scolaires. À partir des moyennes annuelles en mathématiques et en français d'élèves de 3e, on observe des valeurs satisfaisantes : .54 entre le Français et le Raisonnement Verbal (R.V) et .63 entre les mathématiques et le Raisonnement Abstrait (R.A). Une valeur plus élevée (.68) est obtenue, comme c'était déjà le cas dans la version américaine, entre deux scores composites : un indicateur RV + RN et un indicateur composite scolaire (Français + Mathématiques). Toutes ces valeurs indiquent un bon niveau de validité prédictive de la batterie DAT5.

Sur l'échantillon adulte

Une seule étude porte sur cet échantillon. Elle concerne l'analyse interne des intercorrélations. On observe ici des variations un peu moins élevées que celles observées sur l'échantillon scolaire avec des valeurs de .15 (entre Relations Spatiales et Orthographe) à .62 (entre Orthographe et Grammaire), mais on retrouve globalement le même patron de résultats (par exemple des valeurs autour de .55 entre les trois principales épreuves de raisonnement).

La standardisation

➤ La passation

Le matériel DAT 5 se compose d'un manuel (81 pages), de livrets de passation et de grilles de cotation. Il n'y a pas de feuille de passation (excepté pour l'épreuve de *Vitesse de Perception*) : le sujet répond directement sur le cahier de passation.

Le matériel est de type papier-crayon, destiné à une utilisation collective mais bien entendu, comme tout test collectif, une utilisation en individuelle est possible.

Les huit épreuves sont organisées en cinq cahiers : le cahier 1 regroupe les trois aptitudes principales (Raisonnement Verbal, Numérique et Abstrait), les subtests Orthographe et Grammaire sont regroupées dans le cahier 2, les autres aptitudes étant sur des cahiers séparés. Les consignes et les exemples de chaque épreuve figurent au début des cahiers de passation.

Sont indiqués également au sujet le nombre d'exercices (d'items) et le temps de passation. Les temps varient de 6 minutes (VPP) à 20 minutes selon les épreuves.

➤ La cotation

La cotation est très rapide : elle s'effectue à partir de grilles de correction transparentes que l'on superpose aux réponses des sujets. On attribue 1 point pour chaque bonne réponse afin d'obtenir un score brut pour chaque subtest évalué.

Le psychologue peut également additionner le score brut Raisonnement Verbal et le score brut Raisonnement Numérique pour obtenir un score composite (RV + RN). Au total, le sujet peut donc être caractérisé par neuf scores.

➤ Les étalonnages

Les étalonnages comportent onze classes (de 0 à 10). Aucune indication n'est cependant fournie concernant le type d'étalonnage qui a été utilisé. Par déduction on suppose qu'il s'agit d'un étalonnage normalisé en 11 classes, mais c'est au psychologue de le découvrir !

Rappelons qu'il existe deux types d'étalonnages (les étalonnages normalisés et les étalonnages par quantilage), et qu'il est nécessaire de connaître le

type d'étalonnage qui est proposé dans le manuel afin d'éviter des erreurs d'interprétation (voir chapitre 2 de ce livre si nécessaire).

Les auteurs n'indiquent pas non plus les proportions théoriques de sujets de chaque classe, ce qui ne facilite pas la tâche du praticien. Rappelons que dans ce type d'étalonnage la classe centrale (ici la classe 5) regroupe toujours l'effectif les plus important (ici 15,9 %), avec une diminution progressive de la proportion de sujets au fur et à mesure que l'on se déplace vers les classes extrêmes, avec par exemple ici[1] 14,6 % dans la classe 4 (ou 3) mais 3,6 % dans la classe 10 (ou 0).

Les étalonnages portent sur deux types de population, adolescents scolarisés et adultes, avec distinction entre plusieurs niveaux de formation :

- Concernant les étalonnages « adolescents », quatre niveaux scolaires sont disponibles : 3e, 2e, 1ère/Terminale, et enseignement professionnel ;
- Concernant les étalonnages « adultes » : trois niveaux : CAP/BEP, Baccalauréat et étalonnage total.

Pour certains étalonnages « adolescents », (lorsque les différences sont significatives), figurent des étalonnages séparés par sexe. C'est le cas principalement du Raisonnement Spatial ou encore de Vitesse et Précision.

Une fois l'étalonnage sélectionné (il doit être le plus proche possible des caractéristiques du sujet) il ne reste plus qu'à transformer chaque score brut en score étalonné.

Comme dans les autres épreuves, il peut être intéressant de confronter les résultats d'un sujet à plusieurs populations de référence.

L'interprétation des scores de la DAT5

Mise à part la recommandation de tenir compte de l'erreur de mesure[2], les auteurs du manuel de la DAT5 ne fournissent aucune information concernant l'analyse des résultats : ni méthode d'interprétation des résultats, ni étude de cas.

Nous proposons donc ici une procédure d'interprétation des scores en quatre étapes, globalement semblable à celle proposée dans les autres batteries : de la prise en compte des résultats de chaque subtest à

1. Le lecteur pourra se référer au tableau 2.3 du chapitre 2 de ce livre pour connaître précisément les effectifs théoriques de chacune des classes.

2. Les auteurs proposent un exemple de calcul et de prise en compte de l'erreur de mesure (manuel, p. 19)

l'interprétation du profil des résultats en rapport avec les objectifs de l'évaluation.

Propositions de procédure d'interprétation des scores de la DAT

Étape 1 : Analyse de chaque score étalonné

Le psychologue commencera par reporter les notes brutes et les scores étalonnés dans le tableau de profil situé sur la première page du cahier 1. Étonnamment ne figure pas, dans ce tableau, un emplacement pour le score composite RV + RN alors même que le psychologue dispose d'étalonnages pour ce score. Est-ce un oubli ? Le psychologue devra rajouter manuellement une ligne à ce tableau pour y faire figurer cet indicateur.

Pour chaque aptitude évaluée, la performance du sujet peut varier de 0 à 10. Comment interpréter ces valeurs ? Nous proposons de regrouper des scores comparables[1].

Rappelons qu'avec un étalonnage en 11 classes normalisées, un score situé dans la classe centrale (classe 5) correspond à un score moyen. On interprétera de la même manière un score étalonné de 6 (légèrement au-dessus de la moyenne) ainsi qu'un score de 4 (légèrement en dessous de la moyenne). Il faut se rappeler que ces trois classes centrales (4, 5 et 6) regroupent près de 50 % des sujets[2]. Un score situé dans l'une de ces trois classes indique donc que la performance du sujet se situe dans la moyenne de l'échantillon de référence.

Les scores inférieurs pourront être qualifiés de faibles (scores 2 et 3), voire très faibles (scores 0 et 1) et, symétriquement, les scores supérieurs pourront être qualifiés d'élevés (scores 7 et 8), ou très élevés (scores 9 et 10).

Si le psychologue souhaite être plus précis, il peut estimer la proportion de sujets qui obtient une note plus élevée, ou moins élevée, que celle du sujet évalué.

Comment interpréter par exemple un score étalonné de 3 ?

On peut déjà indiquer que c'est un résultat faible, très inférieur à la moyenne. Ensuite on peut estimer la proportion de sujets qui se situe

1. Ce qui est également une manière de tenir compte de l'erreur de mesure.
2. 45,1 % très exactement (14,6 + 15,9 + 14,6 = 45,1 %)

en dessous, et au dessus, de ce score[1] : seulement 15,8 %[2] des sujets obtiennent un score plus faible, 11,6 % (proportion de la classe 3) obtiennent un score comparable, et 72,6 % des sujets obtiennent donc un score supérieur. Ces éléments confirment bien la faiblesse d'un score étalonné de 3.

Bien entendu, la même démarche s'applique pour les scores élevés.

Étape 2 : Analyse du profil

Il faut distinguer ici deux types de comparaison : interindividuelles et intra-individuelle.

Dans une comparaison *interindividuelle* le psychologue sera sensible aux scores qui atteignent ou dépassent un certain seuil, par exemple la note de 7, et qui représentent les points forts du sujet, comparativement aux autres sujets de l'étalonnage.

À l'inverse, les scores égaux ou inférieurs à 3 peuvent être considérés comme des points faibles.

Dans une comparaison *intra-individuelle*, le psychologue sera sensible au tracé du profil plus qu'aux valeurs des scores. Il s'agit ici de repérer les valeurs élevées, et les valeurs faibles, mais en référence cette fois non pas aux scores de l'étalonnage mais aux autres scores du sujet (sa moyenne personnelle). Ainsi, par exemple, pour un sujet qui n'obtient pas globalement de bons résultats dans la comparaison interindividuelle (avec des scores situés par exemple entre 2 et 5 selon les subtests) le psychologue sera attentif au profil et aux scores extrêmes (ici les scores 5 et 2) qui détermineront les points forts, et faibles, du profil du sujet en référence cette fois à l'ensemble de son profil.

Étape 3 : Analyse du score composite RV + RN

Rappelons que ce score composite présente une bonne qualité prédictive par rapport aux acquisitions scolaires. Le psychologue interprétera ce score dans ce sens et dans une comparaison interindividuelle.

Étape 4 : Analyse des relations entre aptitudes évaluées et activités professionnelles

Le psychologue trouvera dans le manuel (p. 5 à 12) des précisions sur ce qui est évalué dans chaque subtests ainsi que des informations

1. Voir tableau 2.3 du chapitre 2.
2. Si on additionne les pourcentages des classes 0, 1 et 2 on obtient 3,6 + 4,5 + 7,7 = 15,8 %

succinctes concernant les liaisons entre ces aptitudes et certaines activités professionnelles. Par exemple, pour le Raisonnement Verbal il est signalé :

> « Ce test peut être utilisé pour aider à prédire le succès scolaire mais aussi le réussite dans certaines disciplines telles que le commerce, le droit, le journalisme, l'enseignement, les sciences. » (manuel, p. 6).

Dans le cadre d'une pratique de conseil, mais dans une moindre mesure, dans le cadre d'une pratique de sélection, il ne s'agit pas, comme nous l'avons déjà signalé à propos des autres batteries, de rechercher une parfaite adéquation entre le profil du sujet et les activités professionnelles exercées et/ou envisagées, mais plutôt d'aider le sujet à prendre conscience des éventuels décalages qui peuvent exister entre son projet professionnel et l'estimation de ses aptitudes. Ces décalages pouvant d'ailleurs faire l'objet d'échanges avec le sujet lors de la procédure de restitution des résultats.

Conclusion sur la DAT5

Directement adaptée d'une épreuve américaine, la batterie DAT5 est une batterie assez complète, particulièrement bien adaptée à des publics scolaires (de niveau 3e à Baccalauréat). Utilisée dans sa totalité elle permet de caractériser chaque sujet selon huit scores d'aptitudes et selon un score composite supplémentaire. Les différents subtests peuvent être utilisés séparément, ce qui offre une grande souplesse d'utilisation. Ses qualités psychométriques sont globalement satisfaisantes, mais on peut regretter le manque d'informations concernant l'interprétation des scores, et tout particulièrement l'absence d'études de cas.

TROISIÈME PARTIE

Utilisation des tests d'intelligence

CHAPITRE
6

De la mesure des performances à l'analyse des stratégies

Sommaire

L'**OBJECTIF** principal des tests, et plus précisément des tests d'aptitude, est de différencier les sujets sur une dimension définie, relativement stable, comme, par exemple, l'aptitude cognitive générale (ou facteur *g*), ou une aptitude spécifique. Cette différenciation, que permet la sensibilité des épreuves, porte sur le niveau de performance observé (le ou les scores) dans une tâche précise (les items du test).

Cette approche *quantitative* de la variabilité interindividuelle représente le principe général de la mesure dans les tests que nous avons présenté dans les chapitres précédents.

Nous aborderons maintenant dans ce chapitre les principaux éléments d'une autre approche des différences individuelles, une approche plus *qualitative* qui peut refléter l'évolution de la psychologie différentielle de ces dernières années. En effet, nous sommes passés d'une analyse de la variabilité interindividuelle en terme d'efficience et de niveau de performance (liée au courant psychométrique et à la méthode des tests) à l'analyse des différences interindividuelles en terme de fonctionnement cognitif. Ce dernier type de différences permettant éventuellement d'expliquer les variations observées en terme d'efficience :

> « Les différences de performances peuvent s'expliquer d'abord, dans une chaîne causale qu'il convient d'exploiter maillon après maillon, par la mise en œuvre de modalités de fonctionnement différentes chez des individus différents confrontés à la même situation. » (Reuchlin, 1990a, p. 15).

Cette approche peut être qualifiée de « cognitivo-différentielle ». Elle présente de nombreux points communs avec, d'une part, l'approche intégrative[1] proposée par d'autres chercheurs (Rozencwajg, 2005 et 2003), et, d'autre part, avec l'approche du diagnostic cognitif[2] (Richard, 1996). L'objectif de cette approche n'est donc plus de différencier les sujets (uniquement) sur leur niveau de performance mais de les caractériser également sur les modalités de réalisation de cette performance, sur la singularité de leur fonctionnement cognitif, et plus précisément, comme nous allons le présenter ensuite, sur leur(s) stratégie(s) de résolution.

1. « *L'approche intégrative consiste à identifier les différents aspects du fonctionnement au sein de tâches complexes par l'identification des stratégies de résolution. Ces stratégies, qualitativement différentes, sont révélatrices de fonctionnements individuels différents.* » (Rozencwajg, 2005, p. 105).

2. « Le diagnostic se rattache à une approche différentielle de l'étude des processus cognitifs. » (Richard, 1996, p. 4).

Les recherches dans ce domaine sont nombreuses (voir par exemple l'ouvrage *Perspectives différentielles en psychologie,* Loarer *et al.* 2008) mais l'écart est important entre résultats de recherches et applications concrètes. Bien que l'on dispose maintenant d'une assez grande variété de situations d'évaluation permettant d'illustrer ce type de différences individuelles, ces situations relèvent le plus souvent d'épreuves expérimentales et on ne dispose pas encore de versions réellement utilisables, c'est-à-dire de tests édités, validés et étalonnés. En effet, le passage est souvent difficile entre les recherches et les pratiques.

> « Dans le domaine de l'intelligence, les méthodes de diagnostic cognitif restent, pour l'instant, du domaine de la recherche », observaient déjà Huteau et Lautrey en 1999a (p. 247).

Pourtant l'élaboration de ce type de test présenterait un grand intérêt :

> « Il semble (...) qu'il y ait beaucoup à attendre d'épreuves – à construire celles-là – qui caractériseraient les individus selon les stratégies qu'ils utilisent. » (Huteau, 1985b, p. 83).

Le logiciel *SAMUEL* (Rozencwacg, Corroyer et Altman, 1999, 2001), que nous présenterons en fin de chapitre, constitue sans aucun doute l'exception qui confirme la règle car il permet d'évaluer les stratégies de résolution des sujets dans une épreuve informatisée de type cubes de Kohs. Avant de présenter ce test, et les études expérimentales sur lesquelles il repose, il nous faut préciser davantage cette notion de stratégie de résolution.

1. La notion de stratégie

> « Lorsqu'on observe plusieurs individus accomplissant la même tâche, on constate souvent en effet qu'ils ne procèdent pas tous de la même façon. On dira qu'ils n'utilisent pas tous la même stratégie. » (Reuchlin, 1997, p. 117).

Les situations de tests d'intelligence peuvent être considérées comme des situations typiques de résolution de problème dans lesquelles il est possible d'observer de telles différences entre individus (Richard et Zamani, 1996). Le test des cubes de Kohs en est un bon exemple, mais on pourrait également citer le test du Passalong (voir par exemple les travaux de Beuscart-Zéphir,

Anceaux, Duhamel et Quentin, 1996 et ceux de Richard et Zamani, 1996) ou encore le test D70 (voir par exemple les travaux de Rémy, 2001a).

Dans ces situations, où l'attention va se déplacer de l'analyse de la performance vers l'analyse de la résolution, l'évaluation portera sur l'activité du sujet pendant toute la passation, et non plus seulement sur ses résultats :

> « L'identification des processus de traitement s'appuie ici sur l'observation en temps réel du déroulement de la conduite du sujet en train de résoudre les items du test. » (Huteau et Lautrey, 1999a, p. 239).

Le niveau de précision de cette analyse peut varier et aura comme conséquence d'apporter quelques nuances à cette définition :

> « Si l'analyse est peu poussée la stratégie pourra être assimilée à un type d'opération, à une opération réalisée plus ou moins efficacement, ou encore, si on met l'accent sur la préparation de l'activité, à une attitude. La stratégie ne caractérise plus vraiment la structure de l'activité mais plutôt son allure générale. » (Huteau, 1985b, p. 71)

Ainsi, selon le niveau d'analyse, la stratégie peut être un type d'opération mentale, une séquence d'opérations (suite d'opérations mentales) ou encore une attitude.

Pour Reuchlin, stratégie et procédure de résolution semblent être synonymes :

> « Des individus différents emploient souvent des procédures (on dit aussi des « stratégies ») différentes pour exécuter la même tâche. » (Reuchlin, 1997, p. 107)

Cette diversité de stratégie, cette redondance de fonctionnement, est directement liée, pour ce même auteur, à la diversité des processus mentaux :

> « Cette diversité des stratégies peut être attribuée à la diversité des processus mentaux qui sous-tendent l'exécution de la tâche. » (Reuchlin, 1997, p. 117)

Cette diversité, aussi bien au niveau des processus mentaux que des stratégies, permet alors à des individus différents d'utiliser des moyens différents dans la résolution d'une même tâche :

> « Des composantes ou processus différents peuvent être mis en œuvre par des stratégies différentes permettant toutes de résoudre le problème posé. » (Reuchlin et Bacher, 1989, p. 136)

Lautrey rapproche également ces deux notions, stratégies et processus, dans la situation de résolution de problème :

> « À l'échelle du temps de la résolution de problèmes, on parle généralement de différences de stratégies pour désigner ces différences dans le choix des processus. » (Lautrey, 1995, p. 8)

Enfin, distinguer les sujets sur des différences de stratégies, sur des différences de processus, c'est les différencier sur des variables qualitatives :

> « Nous réservons l'expression "différences de stratégie" à des différences qualitatives dans la nature des processus mobilisés pour résoudre un même problème. » (Huteau et Lautrey, 1999a, p. 232)

Ces définitions de la notion de stratégie permettent de préciser ce qui va constituer ici l'objet même de l'évaluation. Cette approche, appliquée à la résolution d'items de tests, peut se présenter ainsi : il est possible que des sujets différents, face à un même item, mettent en place des stratégies différentes pour aboutir finalement au même résultat (la bonne réponse).

Ce type d'analyse n'est jamais (ou quasiment jamais) effectué dans les tests disponibles actuellement. Pourtant, bien avant les propositions que nous venons rapidement de présenter (et qui datent des années 1980-1990), des auteurs se sont intéressés à la démarche du sujet, aux procédures de résolution d'items de tests. Par exemple dans les années 1930, Kreutz (1934) va consacrer un article à la problématique de « l'inconstance » des tests. Son objectif est différent de celui exposé dans ce chapitre, il ne s'agit pas pour lui d'étudier finement les démarches de résolution pour en faire un objet d'étude, mais il souhaite les analyser afin de réduire les sources de variations individuelles et ainsi, réduire l'erreur de mesure (suivant l'approche psychométrique classique dominante à cette époque). Certaines de ses réflexions nous apparaissent encore très pertinentes aujourd'hui et semblent annoncer les principes d'une analyse cognitivo-différentielle que d'autres auteurs proposeront plus tardivement. Il note ainsi que :

> « Les tâches même les plus simples en apparence, peuvent être résolues très différemment (...) Le moyen de résoudre la tâche est important à connaître, vu que les comportements différents mobilisent des aptitudes différentes ; par conséquent, si l'on ignore le moyen de résoudre la tâche, il est impossible d'interpréter correctement le résultat. » (Kreutz, 1934, p. 229)

Mais au lieu de s'intéresser plus précisément à cette variabilité il cherche alors à la réduire et propose par exemple de « fixer la façon d'agir, la technique du sujet » (p. 234) de façon à rendre comparable les résultats. Et si certains tests ne permettent pas cette uniformisation de la démarche de résolution il faut « les retirer de la circulation » (p. 235) !

Pour pouvoir fixer cette modalité de résolution il faut commencer par recenser les différentes stratégies possibles :

> « D'après une analyse approfondie du test, connaître tous les moyens possibles de le résoudre. » (p. 235)

Plus de 70 ans plus tard, en relisant ces propositions nous pourrions nous demander si les principes de l'analyse cognitivo-différentielle n'étaient pas déjà énoncés par Kreutz en 1934, même si nous ne reprenons pas à notre compte toutes ses conclusions.

2. Vicariance et affordance

Dans un article de 1978, Maurice Reuchlin présente avec le modèle de la vicariance, un cadre conceptuel visant à expliquer les différences individuelles. Il propose de considérer qu'un même individu dispose de plusieurs processus vicariants (processus pouvant se substituer les uns aux autres) pour élaborer sa réponse et s'adapter à une situation. Chaque sujet disposerait ainsi d'un répertoire[1] intra-individuel de processus vicariants. Le recours à tel ou tel processus pourrait varier selon les individus – certains de ces processus étant plus évocables chez un individu donné – ce qui permettrait d'expliquer les différences interindividuelles observées au niveau de la performance, tous les processus n'étant pas équivalents en terme d'efficacité. Ainsi les individus pourraient être différenciés sur leur hiérarchie d'évocabilité des processus, en raison de la « diversité génétique interindividuelle et la diversité des histoires individuelles » (Reuchlin, 1978, p. 135).

Cette pluralité de processus permet alors au sujet de s'adapter à différentes situations : les processus mis en œuvre par un même sujet pouvant être différents dans des situations différentes. Pour Reuchlin il s'agit d'un modèle probabiliste, qui pourrait être formalisé ainsi : « pour un individu I placé dans une situation S, le processus P a une certaine probabilité d'être évoqué », et qui distingue deux types de paramètres : individuels et situationnels.

Pour les paramètres individuels, Reuchlin envisage l'existence d'une certaine stabilité dans le fonctionnement :

> « Il se trouve, pour des raisons encore mal définies, que chaque individu, même s'il dispose potentiellement de tout le répertoire procédural

1. On trouve aussi dans la littérature le terme « catalogue ».

> caractéristique de son espèce, accorde de façon relativement stable une priorité d'évocation plus ou moins forte à certains de ces processus vicariants. » (Reuchlin, 1990a, p. 20.)

Cette stabilité peut permettre alors d'identifier la stratégie préférentielle du sujet, pour une classe donnée de situations.

Concernant les paramètres situationnels, ils vont agir sur le niveau d'efficacité des processus, et au final, sur le niveau de performance des sujets :

> « Dans une situation déterminée, tous ces processus ne seraient pas également efficaces : ils seraient plus ou moins coûteux pour le même niveau d'adaptation ou auraient des probabilités inégales de conduire à la réussite. » (Reuchlin, 1978, p. 135.)

Ces propositions de Reuchlin vont avoir des retombées importantes dans les recherches en psychologie différentielle : nombre de chercheurs vont ainsi utiliser ce modèle, ou des aspects de ce modèle, pour expliquer des différences inter (et intra) individuelles observées dans différents types de situation. Par exemple, pour Lautrey :

> « Cette notion de vicariance offre un cadre conceptuel intéressant pour rendre compte des différences individuelles qui ont été appelées qualitatives (...) c'est-à-dire des différences tenant à ce que les individus peuvent résoudre un même problème par des processus différents, des stratégies différentes. » (Lautrey, 1999, p. 54)

Ohlmann va s'intéresser tout particulièrement au second type de paramètre avec la notion d'*affordance* (Ohlman, 1991, 1995). Ses recherches concernent plus précisément les effets des caractéristiques de la situation sur les stratégies utilisées par les sujets.

Pour cet auteur, l'observation de différences interindividuelles en terme de stratégies ne peut se réaliser que dans certaines conditions, situations dites à contrainte faible, qui laissent une « marge de liberté » au fonctionnement individuel des sujets. Dans ces situations l'expression de différences interindividuelles stratégitaires est alors possible. Par contre, d'autres situations, à niveau de contrainte plus élevé, ne seraient pas propices à l'expression de ces différences. Chaque situation peut donc être caractérisée par son niveau de contrainte, et c'est ce niveau qui va déterminer les possibilités d'expression de stratégies différentes.

Pour caractériser les situations, Ohlmann propose d'utiliser la notion d'*affordance*, concept proposé par Gibson en 1979 dans le domaine de la perception. Ohlmann nous en propose une première définition en 1990 :

> « C'est une possibilité d'action établie à partir des relations entre un organisme et son environnement. » (Ohlmann, 1990b, p. 425.)

Définition qu'il complète un peu plus tard :

> « Le concept d'*affordance* pourrait se définir comme la perception d'une utilité. Autrement dit, pour un organisme inséré dans un environnement, c'est la possibilité d'effectuer une action finalisée. » (Ohlmann, 1991, p. 212.)

Ainsi les *affordances* perçues/crées par le sujet vont dépendre des caractéristiques de la situation. Mais, à situation identique, elles peuvent différer en fonction des sujets. Ohlmann propose donc une approche différentielle de la notion d'*affordance* :

> « On peut suggérer aussi que des individus différents – quoique d'une espèce identique – tireront d'un environnement identique des *affordances* différentes. Autrement dit, selon l'item qui est placé en tête de catalogue individuel de processus, les propriétés utiles du milieu seront perçues différemment. » (Ohlmann, 1991, p. 214-215.)

Les recherches d'Ohlmann vont alors porter principalement sur ces interactions individu-situation dans la cognition spatiale, et plus précisément sur l'analyse des conduites posturales et des situations de conflit vision/posture (Ohlmann, 1990a, 1990b, 1991, 1995, 2000).

3. Comment identifier les stratégies ?

Identifier la stratégie de résolution d'un sujet dans une tâche donnée nécessite de disposer préalablement de données fiables concernant les éléments suivants :

- Connaître toutes les possibilités de résolution (lister l'ensemble des stratégies possibles), ;
- Sélectionner les indicateurs pertinents pour chacune de ces stratégies ;
- Élaborer un dispositif adapté de recueil de données.

Nous avons vu précédemment que, dès 1934, Kreutz proposait d'observer les différents modes de résolution des items de test. Il envisageait alors trois possibilités méthodologiques :

1° Recourir à l'introspection ;

2° Observer les conduites en cours de résolution ;
3° Procéder à une « analyse raisonnée du test » (on pourrait dire aujourd'hui une analyse *cognitive* du test) afin de repérer les différentes possibilités de résolution (Kreutz, 1934, p. 238).

Nous pouvons retrouver des liens entre ces trois pistes de Kreutz et différentes méthodologies utilisées actuellement dans les recherches. Nous ne présenterons ici que trois exemples, l'un, l'analyse des temps de réponse, car c'est sans doute historiquement l'une des premières méthodes employées, les deux autres, l'analyse dynamique de la résolution et la création d'un matériel spécifique, car ils sont directement liés à l'exemple de l'épreuve SAMUEL que nous présenterons ensuite.

L'analyse de la structure des temps de résolution

Cette méthode chronométrique a été largement utilisée dès les premières recherches de psychologie cognitive différentielle à partir du postulat suivant : des traitements (stratégies) différents vont se différencier par des patrons différents de temps de traitement. La validation des modèles théoriques des stratégies (modèles stratégitaires) reposera sur l'analyse des temps de préparation et/ou de résolution. Ainsi on a montré, par exemple dans des tâches de rotation mentale, que l'utilisation d'une stratégie analogique (imagée) pouvait se repérer par une liaison linéaire entre le temps de résolution et la valeur de l'angle de rotation de la figure tandis que dans le cas de l'utilisation de la stratégie propositionnelle, le temps de résolution n'est pas en relation directe avec l'angle de rotation (voir par exemple Gilles, 1991 et 1993).

Mais le caractère univoque de la signification des temps de réponse a été remis en cause en raison de deux critiques principales (Marquer et Pereira, 1987, 1990) :

- À un même patron temporel peuvent correspondre des traitements différents ;
- Des patrons identiques peuvent traduire des traitements différents.

À partir d'analyses de résultats de recherches, ces auteurs nous indiquent bien que ce type d'indicateur n'est pas toujours valide.

L'analyse dynamique de la résolution

Cette méthode, qui consiste à suivre pas à pas la démarche du sujet dans la résolution du problème, a été très utilisée pour l'étude de la résolution d'items de tests. Plusieurs méthodes de recueil des données sont utilisables :

– l'étude des mouvements oculaires ;
– l'enregistrement vidéo ;
– l'automatisation du recueil des variables.

Nous présentons quelques recherches qui illustrent ces différentes méthodologies.

➤ L'étude des mouvements oculaires

Dans une étude portant sur la résolution des *Progressives Matrices* de Raven (Carpenter, Just et Shell, 1990), les auteurs vont analyser les procédures de résolution des sujets à partir des données suivantes : temps de résolution, erreurs, verbalisation des règles trouvées en cours de résolution et mouvements oculaires pendant la résolution.

Cette dernière variable leur permet d'analyser très finement les regards des sujets, et la dynamique de ces regards, afin, par exemple, de savoir quelles sont les matrices les plus regardées, quelles sont celles qui subissent le plus d'allers-retours visuels... et de comprendre alors la dynamique de résolution du sujet dans la découverte des règles de transformation des éléments de la matrice. Les auteurs montrent ainsi que les sujets décomposent souvent le problème global en plusieurs problèmes plus élémentaires (comme la recherche de la règle de progression entre deux éléments). Les auteurs élaborent alors un programme informatique (*Fairaven*) censé simuler la procédure de résolution des sujets les moins performants, puis un second programme (*Bettaraven*) adapté à la procédure de résolution des sujets les plus performants. On peut considérer ces deux programmes comme représentatifs des deux types de stratégies de résolution, la seconde étant la plus efficace.

➤ L'enregistrement vidéo des passations

Une observation directe de la conduite du sujet en temps réel est assez difficile à réaliser, même à l'aide de grille d'observation, car les informations à prendre en compte sont très nombreuses. Un enregistrement vidéo permet

le recueil de toutes les observations d'ordre comportemental qui pourront ensuite être analysées et codées afin de fournir des indicateurs caractéristiques des stratégies utilisées par les sujets. Une telle méthodologie a été utilisée, par exemple, dans la résolution des cubes de Kohs (Beuscart-Zéphir et Beuscart, 1988 ; Rozencwajg, 1991). Ces recherches confirment l'existence de plusieurs stratégies dans cette épreuve (nous développerons plus loin ces études).

➤ L'automatisation du recueil des observables

L'enregistrement vidéo des passations, s'il présente de nombreux avantages, comporte également des inconvénients, en particulier une lourdeur méthodologique (caméras, codage des bandes vidéos...) qui limitent son utilisation. Des chercheurs ont alors envisagé d'automatiser le recueil des données, soit en conservant la situation mais en utilisant un dispositif électronique qui enregistre les données (voir par exemple Beuscart-Zéphir et Beuscart, 1989, et Beuscart-Zéphir *et al.*, 1996, sur l'épreuve du *passalong*), soit en informatisant la tâche (voir par exemple Rozencwajg, Corroyer et Altman, 1999, et Houssemand, 1999a et 1999b, sur l'épreuve des cubes de Kohs ; Richard et Zamani, 1996, sur l'épreuve du *passalong*).

Dans le premier cas, la situation est strictement la même pour le sujet (en comparaison des modalités de passation classique de l'épreuve), par contre, dans le second cas, l'épreuve est profondément modifiée de part l'informatisation (transformation de la relation sujet/tâche du test par la médiation de l'ordinateur, passage en deux dimensions...).

Ces procédures d'automatisation du recueil des données ne consistent pas uniquement à automatiser la passation et la correction (comme c'est simplement le cas dans la grande majorité des tests informatisés) mais permettent de recueillir de manière automatique plusieurs types d'indicateurs stratégitaires pertinents pour caractériser les différentes stratégies (temps de résolution, écart temporel entre deux actions, ordre des actions...). À partir de ces indicateurs il est possible d'élaborer des modèles théoriques des stratégies, voire de réaliser un diagnostic automatique de la stratégie utilisée par le sujet (voir plus loin l'exemple du logiciel SAMUEL).

La création d'un matériel spécifique

Une approche originale pour repérer la stratégie des sujets a été proposée et mise en œuvre par Rémy (2001a et 2001b). À la suite des premiers travaux de Dickes et Martin sur le test D70 (Dickes et Martin, 1998), dans lesquels ces auteurs proposent une catégorisation des items des tests de type dominos (voir chapitre 4), Rémy va élaborer une épreuve originale reposant sur des items pouvant être résolus selon deux logiques différentes : soit par une stratégie spatiale, soit par une stratégie numérique. Et ce qui fait l'originalité de cette approche c'est que la réponse du sujet à ces items (nommés items « équipotents[1] ») sera différente selon la stratégie qu'il a utilisée.

Autrement dit, il y a pour chaque item deux bonnes réponses, chacune témoignant de l'utilisation d'une stratégie définie. L'inférence de la stratégie utilisée par le sujet présente donc ici toutes les garanties d'objectivité car la réponse est univoque et dépend entièrement (en cas de bonne réponse) de la procédure utilisée. À notre connaissance il n'existe pas de version utilisable de cette épreuve.

Un second exemple de création de matériel spécifique porte sur l'épreuve des cubes avec la création d'un logiciel, SAMUEL, dont l'objectif principal consiste à évaluer la stratégie utilisée par le sujet. Ce sera l'objet de la partie suivante.

4. *De l'analyse des stratégies dans l'épreuve des cubes de Kohs au logiciel SAMUEL*

L'analyse des stratégies dans l'épreuve des cubes

L'épreuve des cubes de Kohs est une épreuve très connue des psychologues. Rappelons qu'elle a donné lieu à de multiples versions et à de nombreuses recherches (Chartier, 2002a). Nous nous intéresserons tout particulièrement dans cette partie aux données concernant l'analyse des stratégies de résolution.

Wechsler, qui avait introduit un subtest cubes dès les premières versions de ses échelles d'intelligence, notait dès 1944 (Wechsler, 1956 pour la

1. Items permettant l'utilisation des deux stratégies (soit l'une, soit l'autre) pour arriver à la bonne réponse.

traduction française) qu'il existait une liaison entre la façon dont les sujets se représentaient le modèle et le niveau de réussite à l'épreuve :

> « Assez curieusement, les individus réussissant le mieux ce test ne sont pas ceux qui voient, ou tout au moins suivent, le modèle comme un tout, mais ceux qui sont capables de le fractionner en petites portions. » (Wechsler, 1956, p. 113)

Et cette différence interindividuelle dans la représentation mentale du modèle peut avoir une incidence sur la méthode utilisée par le sujet :

> « Nous avons déjà mentionné la question des différences dans la méthode pouvant être employée pour faire les dessins, à savoir, suivre la figure ou au contraire la décomposer en ses éléments » (Wechsler, 1956, p. 114).

Wechsler distinguait ainsi deux grands types de résolution, que l'on retrouvera ensuite dans les recherches ultérieures :

– soit « *suivre l'image* », stratégie qualifiée ensuite de *globale* ;
– soit « *séparer le dessin en ses parties constituantes* » (p. 113), stratégie appelée par la suite *analytique*.

Cette variabilité possible dans la résolution de la tâche explique que, pour Wechsler :

> « Le test de cubes colorés est non seulement un excellent test d'intelligence générale, mais un de ceux qui se prêtent admirablement à l'analyse qualitative. » (Wechsler, 1956, p. 114)

Ces premières observations de Wechsler rejoignent celles de Goldstein et Scheerer qui distinguaient également, dans la même période, deux approches possibles dans la résolution de cette tâche : une approche concrète et une approche abstraite (Goldstein et Scheerer, 1941).

Ces deux méthodes de résolution, ces deux stratégies, sont à rapprocher des deux principaux types de traitement de l'information qui ont été proposés plus tardivement par les recherches en psychologie cognitive : un traitement global et un traitement analytique.

On retrouve d'ailleurs cette distinction global/analytique dans le manuel du Kohs, diffusé par les éditions EAP[1] :

> « L'observation du comportement permet d'apprécier la qualité de la réussite ou de l'échec, de caractériser le procédé de construction, qui reflète

1. L'épreuve de Kohs a été diffusée à la fois par les éditions EAP et par les éditions ECPA, ce qui explique l'existence de deux manuels pour ce test.

> le niveau de compréhension ; le sujet : qui structure systématiquement le modèle : processus analytique, qui fractionne le modèle en suivant le périmètre, les angles, les motifs, etc : processus semi-analytique, qui procède par tâtonnements, soit systématiques, soit empiriques. » (EAP, 1978, p. 36).

Ces premières propositions des années 1950-1970 confirment bien l'existence, au moins dans cette épreuve, d'une possibilité de variabilité interindividuelle dans la procédure de résolution utilisée par les sujets. Pour le praticien qui s'intéresse à ce type de différences, l'un des problèmes sera alors de pouvoir repérer la stratégie utilisée par le sujet. Une réponse va être fournie par Bonnardel en 1953, pour le test B101 (l'une des versions de l'épreuve de cubes), avec la présentation d'une grille d'observation de la conduite du sujet en cours d'épreuve. Cette grille distingue cinq niveaux de structuration : d'une « analyse par rangée ou colonnes » (catégorisé en A+), à des « essais empiriques, désordonnés » (catégorisé C-).

Si nous reprenons les deux types de constructions proposés par Wechsler (voir plus haut), nous pouvons rapprocher la structuration de type A+ de la stratégie analytique (« séparer le dessin en ses parties élémentaires »), et la structuration de type B de la stratégie globale (« suivre l'image »). Malgré son intérêt, cette grille de Bonnardel ne semble pas avoir été souvent utilisée[1] ni dans les pratiques évaluatives, ni dans les recherches.

Toutes ces approches vont être reprises dans les années 1980 par différentes recherches relevant de psychologie cognitive et/ou différentielle qui vérifient l'existence de ces deux stratégies :

- Une stratégie *analytique* (la plus performante) dans laquelle le sujet procède à une segmentation mentale du modèle en ses composantes élémentaires (les différents cubes), puis identifie la face du premier cube et son orientation avant de le disposer sur l'aire de construction, et continue ainsi sa construction, cube après cube ;
- Une stratégie *globale*[2] dans laquelle le sujet est plus sensible à l'aspect *gestalt* du modèle, et ne parvient pas, ou très difficilement, à opérer cette segmentation mentale du modèle, et procède alors, dans sa construction,

1. Nous trouvons peu de références à cette grille dans la littérature, mise à part une présentation dans l'article de Rozencwajg et Huteau (1996).

2. La plupart des auteurs de langue anglaise utilisent le terme *synthétique* pour désigner cette stratégie, mais nous utiliserons ce terme de stratégie *globale* pour deux raisons : d'une part cette stratégie correspond à la stratégie globale décrite par ailleurs, d'autre part, Rozencwajg a proposé une stratégie spécifique qu'elle a qualifiée de « synthétique » et qui ne correspond pas à cette stratégie globale (voir plus loin).

plutôt par tâtonnements, par essais et erreurs afin de reconstituer la forme globale perçue.

Nous ne reprendrons ici que les principaux résultats des recherches.

Jones et Torgesen (1981) se sont intéressés à l'évolution des stratégies avec l'âge à partir de l'hypothèse suivante : les enfants plus âgés devraient utiliser préférentiellement la stratégie la plus performante, la stratégie analytique. À partir de passations vidéoscopées, ils analysent finement les séquences de placement des cubes (indicateurs de la stratégie utilisée) mais ne constatent pas de différences entre les enfants de différents groupes d'âge.

Schorr, Bower et Kiernan (1982) observent cette variabilité sur une population d'adultes : une stratégie analytique, dans laquelle le sujet procède à une décomposition mentale du modèle, et une stratégie globale, dans laquelle le sujet cherche à reproduire le forme globale. Les auteurs montrent l'existence d'une liaison entre le nombre d'arêtes visibles et la stratégie analytique : plus ce nombre est important, plus la décomposition mentale est facilitée par le modèle, ce qui favorise l'utilisation de la stratégie analytique[1]. Pour les auteurs, la stratégie analytique est, d'une part, la plus employée dans leur échantillon, d'autre part, la plus performante, car la plus rapide.

Spelberg (1987) mène une expérience sur un large échantillon d'enfants (770 enfants de 6 à 16 ans) et trouve également un lien entre le nombre d'arêtes visibles, la rapidité d'exécution et la stratégie analytique. Il suggère également, à la suite de Jones et Torgesen, que le choix de l'une ou l'autre des deux stratégies dépendrait plus de la nature du stimulus que des préférences personnelles du sujet.

Beuscart-Zéphir et Beuscart (1988), dans le cadre général d'analyse cognitive des situations de résolution de problèmes, proposent une formalisation de la tâche des cubes en termes de buts et de sous-buts :

> « L'une des formalisations possibles est la suivante :
> but final : reconstitution du modèle avec des cubes différents (ou avec des cubes identiques orientés différemment). Pour réaliser ce but final, il faut savoir de combien de cubes est constitué le modèle.
> 1r sous-but : identifier le nombre de cubes. Il faut également savoir quel type de cubes mettre en chaque position.
> 2e sous-but : identifier les n cubes, rouge (r) ou blanc (b) ou mixte (m). Il faut enfin, quand c'est nécessaire (faces mixtes), repérer l'orientation du cube.

1. Nous pouvons relier cette observation aux propositions de Ohlmann, exposées plus haut, avec la notion d'*affordance*.

> 3e sous-but : identifier l'orientation des cubes mixtes (m).
> Une fois que les trois sous-buts sont atteints, le but final est directement réalisable. Si le sujet a procédé à l'analyse in extenso, il est capable de mettre correctement en position chaque cube. » (Beuscart-Zéphir *et al.*, 1988, p. 37.)

Les deux stratégies classiques dans cette tâche se différencient alors de la manière suivante : la stratégie analytique, la plus performante, est caractérisée par l'identification des buts et sous-buts alors que pour la stratégie globale « seul le but final est identifié. Le sujet s'engage directement dans le processus de reconstruction. Il assemble les cubes vraisemblablement en utilisant des indices perceptifs ». Lorsque les items deviennent complexes « on retrouve alors la description d'une stratégie par « essais et erreurs » ». (Beuscart-Zéphir *et al.*, 1988, p. 37).

On retrouve bien ici les opérations fondamentales de décomposition mentale du modèle en cubes élémentaires qui caractérisent la stratégie analytique. À partir de passations vidéoscopées les auteurs vont analyser les procédures de construction utilisées par les sujets. Ils distinguent ainsi trois types de construction :

- Une procédure dans laquelle les sujets positionnent directement les faces correctes des cubes au bon endroit (donc peu de manipulation et rapidité d'exécution) qui peut correspondre soit à une stratégie globale réussie, soit à une stratégie analytique ;
- Une procédure comportant plus de manipulations, plus longue, avec des corrections, qui semble correspondre à une stratégie analytique réalisée avec difficulté ;
- Une procédure comprenant de nombreux essais d'assemblages de cubes, dans laquelle le sujet ne cherche pas (ou n'arrive pas) à élaborer une représentation mentale du modèle, procédure qui correspond à une stratégie globale.

Enfin, Rozencwajg (1991) et Rozencwajg et Huteau (1996) vont confirmer, s'il en était besoin, l'existence de ces deux stratégies mais surtout vont identifier l'existence d'une troisième stratégie, stratégie qu'ils nomment « synthétique », qui serait une combinaison des stratégies globale et analytique. Cette stratégie consiste à analyser le modèle en motifs géométriques ou « *gestalts* » (comme par exemple un triangle rouge composé de deux cubes bicolores), motifs que le sujet peut éventuellement retrouver sur plusieurs modèles. Dans sa construction le sujet s'appuie alors sur cette

représentation mentale et reproduit le modèle préférentiellement à partir de ces motifs géométriques. La figure 6.1 présente ces motifs géométriques.

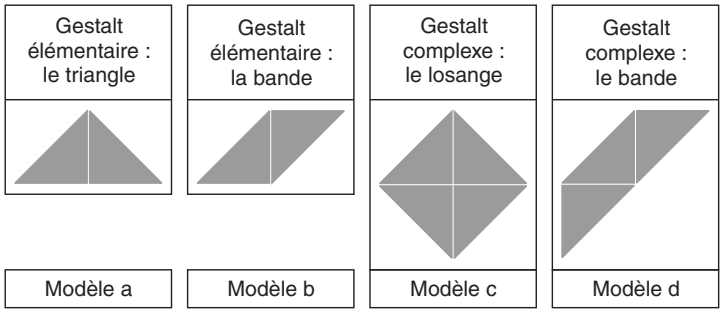

Figure 6.1
Les formes géométriques élémentaires (gestalts) identifiées par Rozencwajg (d'après Rozencwajg, 2005, figure 10, p. 145) **reproduit avec l'aimable autorisation de l'auteur.**

Ces formes géométriques peuvent comporter de deux à quatre faces de cubes : par exemple le triangle est composé de deux cubes, la bande est composée de trois cubes.

On peut remarquer que cette procédure de construction par « motif » avait déjà été proposée en 1978 dans le manuel des EAP et qualifiée alors de *« semi-analytique »* (EAP, 1978, p. 36).

Dans cette stratégie *synthétique*, le sujet procède à une autre forme de segmentation mentale, il ne décomposerait plus (ou plus seulement) le modèle en cubes élémentaires (comme dans la stratégie analytique) mais en groupements de cubes formant une forme géométrique particulière (triangle de 2 cubes, losange de 4 cubes, bande de 3 cubes...).

Pour les auteurs, cette stratégie est plus performante que la stratégie analytique car « elle permettrait au sujet de ne pas traiter tous les cubes isolément et d'avoir de ce fait moins d'informations à prendre en compte » » (Rozencwajg et Huteau, 1996, p. 59).

À la suite de passations vidéoscopées quatre principaux indices ont été retenus pour caractériser ces trois stratégies :

– placement des cubes,
– fréquence du contrôle visuel,
– qualité des séquences
– qualité de l'anticipation.

Les trois stratégies identifiées (globale, analytique et synthétique) se différencient sur ces indices ainsi que sur leur niveau d'efficacité : les stratégies analytique et synthétique étant les plus performantes (avec une légère supériorité de la stratégie synthétique). Les deux principales différences entre les deux stratégies les plus performantes concernent d'une part, la fréquence des regards (fréquence plus faible pour la stratégie synthétique) et, d'autre part, l'ordre de placement des cubes : dans la stratégie analytique le sujet procède préférentiellement par un placement en ligne ou en colonne, cube par cube, dans la stratégie synthétique le sujet privilégie un ordre de placement des cubes selon les gestalts (les motifs géométriques). Les auteurs observent également une forte stabilité intra-individuelle de la stratégie utilisée par le sujet et ils catégorisent alors les sujets sur leur stratégie dominante (ou préférentielle).

Les résultats de Rozencwajg montrent également une liaison entre stratégie et style cognitif D.I.C[1] : les sujets synthétiques et analytiques, sont plus proches du pôle d'indépendance à l'égard du champ, alors que les sujets utilisant la stratégie globale sont plus proches du pôle dépendance.

Les propositions de Rozencwajg sur l'existence de la stratégie synthétique vont être confirmées par d'autres chercheurs (voir par exemple : Houssemand, 1999 ; Chartier, 1999 ; Vrignaud et Chartier, 2003). Certains s'interrogent néanmoins sur le niveau de stabilité intra-individuelle des stratégies observées et sur la possibilité de différences interindividuelles en terme de flexibilité (voir sur cet aspect Chartier, 1999).

Enfin, il faut signaler l'hypothèse d'une quatrième stratégie : la stratégie de *répétition*. Dans un important travail consacré à l'analyse des procédures de résolution dans une épreuve informatisée de cubes, Houssemand va montrer qu'il existe une méthode générale de résolution des modèles à 9 cubes, quels que soient les sujets et quels que soient les modèles :

> « Un placement des faces selon les lignes ou selon les colonnes » (Houssemand, 1999a, p. 228).

Cette méthode correspond à la stratégie analytique souvent décrite dans les recherches. Mais cette stratégie *générale* va lui servir de référence pour repérer d'autres stratégies, plus *spécifiques* (Houssemand, 1999b). Houssemand distingue ainsi la stratégie de « répétition », qui consisterait à repérer des

1. Dépendance/Indépendance à l'égard du champ perceptif (pour une présentation voir Huteau, 2002).

faces identiques de cubes (et orientées de la même façon) présentes dans un même modèle, et à les placer prioritairement à la suite :

> « Il existe un mode de résolution particulier, nommé stratégie de répétition, consistant en un placement prioritaire des faces répétées dans les configurations. » (Houssemand, 1999a, p.272).

Cette répétition de faces, critère de redondance intra-figurale (caractéristique descriptive des modèles de cubes déjà étudiée dans le cadre de l'analyse de la difficulté de la tâche) serait alors support d'un mode de résolution spécifique, éventuellement automatisé. Les résultats indiquent aussi que cette stratégie de répétition n'est pas observée chez tous les sujets et que son utilisation n'est pas liée aux aptitudes cognitives. Néanmoins, l'existence de cette quatrième stratégie mériterait d'être confirmée par d'autres recherches.

Le logiciel SAMUEL de Rozencwajg

Le logiciel *SAMUEL* est présenté comme une épreuve permettant d'établir un diagnostic cognitif[1] à partir d'une version informatisée de la tâche des cubes de Kohs :

> « SAMUEL est une version cognitive du test des cubes de KOHS. Il opérationnalise ce que les psychologues font depuis fort longtemps dans leur pratique des cubes de Kohs, c'est-à-dire un **diagnostic cognitif** qui s'appuie sur des indices cliniques qui vont bien au-delà du temps de résolution et de la performance » (Rozencwajg *et al.,* 1999, p. 11)

Un psychologue même le plus compétent, ne peut prendre en compte qu'un nombre limité d'indices cliniques lors de la passation d'une épreuve. C'est tout l'intérêt d'avoir élaboré ce dispositif informatisé qui permet le recueil et l'analyse d'un grand nombre de variables (voir plus loin). Ainsi,

> « Samuel fournit donc un exemple, assez rare, d'utilisation de l'informatique pour faire d'un test ancien un test vraiment nouveau apportant des informations que le test ancien ne fournissait pas. » (Huteau, introduction du manuel de *SAMUEL*, p. *II*)

C'est un exemple, peut-être le seul, d'épreuve utilisable par un praticien (c'est-à-dire éditée et étalonnée, avec des conditions de standardisation

1. Voir le sous titre du manuel et du test « Samuel. Diagnostic du fonctionnement cognitif » (Rozencwajg *et al.*, 1999).

définies...), et qui a été élaborée spécifiquement pour mettre en évidence des différences interindividuelles qualitatives (les stratégies de résolution).

Signalons que sur le site de l'éditeur de *SAMUEL* (www.delta-expert.com) il est possible de télécharger une version de démonstration de ce logiciel.

➤ Principe général de SAMUEL[1]

Ce test est directement lié aux résultats des recherches de Rozencwajg (voir plus haut) et a été élaboré dans l'objectif de « déterminer automatiquement les stratégies de résolution d'une tâche de performance cognitive » (Rozencwajg *et al.*, 1999, p. 1). La procédure d'évaluation est totalement informatisée : de la présentation des modèles aux calculs des différents indicateurs de la performance du sujet.

La figure 6.2 présente la situation du test *SAMUEL*.

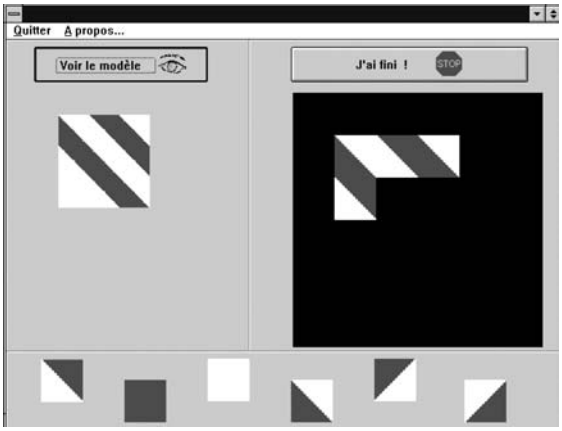

Figure 6.2
***Présentation de la situation du test* SAMUEL *(d'après Rozencwajg, 2005, figure 11, p. 148)* reproduit avec l'aimable autorisation de l'auteur.**

La figure 6.2 représente un modèle en cours de construction. L'écran de *SAMUEL* est composé de trois parties :

– en haut à gauche, apparaît (sur demande du sujet[2]) le modèle à reproduire ;

1. Le nom de ce test est un hommage à *Samuel* Kohs.

2. Le sujet doit cliquer sur l'icône « voir le modèle » pour que le modèle apparaisse. Le modèle reste affiché jusqu'au moment où le sujet clique sur un cube. On mesure ainsi l'un des indicateurs stratégitaires : la fréquence des regards vers le modèle.

– à droite (sur fond noir) figure l'aire de construction ;
– en bas, sont disposées les faces des cubes.

À l'aide de la souris, le sujet doit sélectionner l'une des faces puis la déplacer jusqu'à l'aire de construction. Après quelques items de familiarisation, le sujet doit reproduire quatre modèles (à 9 carrés) qui comportent des formes géométriques identifiées comme *gestalts* par les auteurs.

Bien que les caractéristiques de la tâche soient profondément modifiées par la procédure d'informatisation (passage d'un univers réel à trois dimensions à un univers à deux dimensions, passage des cubes réels en carrés virtuels, contrainte de non rotation des carrés, obligation de manipuler un seul carré à la fois, affichage du modèle sur demande du sujet...), les auteurs reprennent les caractéristiques des trois stratégies de résolution précédemment observées dans leurs recherches[1]. Ainsi, à partir de modélisations théoriques, les trois stratégies (globale, analytique et synthétique) sont alors « identifiées à partir de plusieurs indices de comportement (segmentation, anticipation, fréquence des regards, ordre de placement des cubes par *gestalts*, ordre de placement des cubes par lignes/colonnes) » (Rozencwajg *et al.*, 1999, p. 13).

Le tableau 6.1 présente les valeurs théoriques des indicateurs pour les trois stratégies.

Tableau 6.1
Profils théoriques des trois stratégies (d'après Rozencwajg et al., 1999, p. 41).

	Indices stratégiques de *SAMUEL*				
	Segmentation	Anticipation	Fréquence des regards	Ordre de construction par *gestalts*	Ordre de construction linéaire
Stratégie Synthétique	1	1	0.20	1	0
Stratégie Analytique	1	1	1	0.40	1
Stratégie Globale	0.50	0.50	0.50	0.50	0

L'indice de segmentation porte sur la qualité des séquences (correction des erreurs), l'indice d'anticipation affine l'indice de segmentation et concerne plus particulièrement la proportion de cubes placés correctement du premier

1. Bien qu'il n'existe pas, à notre connaissance, d'études comparatives sur l'utilisation des mêmes stratégies, par les mêmes sujets, dans les deux situations.

coup. Par construction, ces indicateurs varient de 0 à 1 : plus la valeur est proche de 1, plus la construction est de bonne qualité.

La fréquence des regards s'obtient en rapportant le nombre de regards du modèle au nombre d'actions (saisie/repose d'un cube).

Enfin, les deux autres indices sont relatifs au type de placement effectué par le sujet : l'ordre linéaire correspond à un placement effectué en suivant les lignes ou les colonnes, l'ordre par *gestalt* correspond à un placement effectué en suivant les formes géométriques. Plus la valeur est proche de 1, plus le placement observé correspond au placement attendu.

Le logiciel va prendre en compte les valeurs de chaque indicateur pour calculer le profil personnel de chaque sujet. Ce profil sera comparé aux profils théoriques afin de catégoriser le sujet « par la stratégie dont il est le plus proche » (Rozencwajg *et al.*, 1999, p. 42). Il est possible d'obtenir la stratégie utilisée par un sujet sur chacun des items ainsi que sa stratégie préférentielle[1].

Sur leur échantillon, les auteurs retrouvent bien les trois stratégies postulées. Ils observent que la stratégie globale est la plus utilisée (dans 42 % des items), puis la stratégie synthétique (33 %) et enfin, la stratégie analytique (25 %). Cette répartition évolue avec l'âge des sujets dans le sens d'une diminution progressive de l'utilisation de la stratégie globale au profit de la stratégie synthétique.

Ils observent également, au niveau du groupe, des différences de stratégies en fonction des items : les items avec losange sont par exemple plus souvent résolus avec la stratégie synthétique[2].

➤ Indicateurs psychométriques de SAMUEL

Le manuel de *SAMUEL* (Rozencwajg *et al.*, 1999) comporte 116 pages (annexes comprises) et comporte un certain nombre d'informations concernant l'installation du logiciel, le cadre théorique de référence, les qualités psychométriques... Le lecteur y trouvera également des études de cas.

Concernant tout particulièrement les qualités psychométriques, sont présentées des données concernant la fidélité de l'épreuve, ainsi que sa validité.

1. Stratégie sur laquelle le profil du sujet présente la distance la plus faible.
2. On retrouve ici la notion d'*affordance* proposée par Ohlmann.

Par une méthode test/retest la fidélité des indicateurs de *SAMUEL* a été estimée : les corrélations varient de .56 à .79 selon les indicateurs (d'après le tableau 62, p. 63 du manuel). Certaines valeurs semblent un peu faibles, en particulier celles concernant l'ordre de placement des cubes (.56 à .64).

Sur une partie de l'échantillon d'étalonnage (50 élèves de niveau de 5e de collège), qui était confrontée à un ensemble de tests, les auteurs observent les résultats suivants concernant la validité de *SAMUEL* :

- Sur un test spatial (subtest des PMA de Thurstone), comme sur un test de facteur *g* (matrices de Raven) les sujets synthétiques[1] obtiennent, en moyenne, un meilleur score, devant les sujets analytiques puis les sujets globaux ;
- Sur une épreuve de DIC[2] (GEFT) ce sont les sujets analytiques qui obtiennent les meilleurs résultats, puis les sujets synthétiques et enfin les sujets globaux.

Ce dernier résultat est plutôt inattendu car les auteurs pensaient retrouver un lien plus important entre dépendance/champ et stratégie synthétique (en référence aux premiers résultats de Rozencwajg et Huteau, 1996). Des études supplémentaires seraient à mener pour éclaircir ce point.

Rappelons que ces résultats ont été observés sur un échantillon spécifique (50 élèves de collège) et qu'il serait souhaitable de pouvoir disposer de résultats portant sur d'autres populations pour pouvoir généraliser avec confiance ces indications.

Pour la passation il faut utiliser le logiciel *SAMUEL* –Test.

La passation

La passation comporte deux phases :

– une phase de familiarisation avec le dispositif informatique, dans laquelle le sujet doit reproduire des bonhommes et des fleurs ;
– une phase de test, qui comporte six modèles à 4 carrés, et quatre modèles à 9 carrés. Rappelons que le diagnostic concernant les stratégies n'est effectué que sur les modèles à 9 carrés.

Le temps n'est pas limité[3] (bien que le temps de résolution soit l'une des variables prises en compte) et les auteurs estiment à 15 à 20 minutes environ le temps de passation.

1. Les élèves ont été catégorisés dans leur stratégie dominante.
2. Dépendance/Indépendance à l'égard du champ.
3. Il n'y a pas de critères d'arrêt.

La cotation

La cotation est effectuée par le logiciel. Plusieurs variables sont mesurées pour chaque item : réussite, temps de résolution, fréquence des regards, temps de regard total, temps de regard moyen, indices stratégitaires (segmentation, anticipation, ordre de placement). À partir des modèles théoriques des trois stratégies (voir plus haut) le sujet est catégorisé pour chaque item dans la stratégie la plus proche de son profil. Le logiciel détermine également, sur l'ensemble des 4 items, la stratégie dominante (ou préférentielle) de chaque sujet.

Les étalonnages

Le manuel comporte des indications sur six classes d'âge : 9, 11, 13, 15, 17 et 25 ans. Pour chacune de ces classes on dispose de données descriptives concernant les variables précitées (moyenne, écart-type, mini, maxi).

Pour les résultats étalonnés, ils sont organisés selon deux possibilités : par groupe d'âge (les six classes) et par stratégies. Il s'agit d'étalonnages d'effectifs égaux (cinq classes comportant chacune 20 % de l'échantillon).

➤ Interprétation des résultats

Le manuel comporte plusieurs types d'informations utiles au praticien : un guide d'interprétation des résultats (p. 105) et des études de cas (p. 77 à 87). L'interprétation portera principalement sur la résolution des quatre modèles à 9 carrés.

Le principe général d'interprétation se divise en deux phases :

- L'analyse de la variabilité interindividuelle, dans laquelle le sujet sera situé sur sa stratégie dominante et sur ses résultats obtenus sur les différentes variables mesurées, comparativement aux sujets du même groupe d'âge ;
- L'analyse de la variabilité intra-individuelle, dans laquelle l'attention sera portée ici sur les évolutions éventuelles de la stratégie utilisée par le sujet à travers les quatre items (stabilité ou flexibilité ; apprentissage en cours d'épreuve...).

Les études de cas du manuel reposent sur l'analyse des protocoles de 12 sujets, avec confrontation des indices de *SAMUEL* avec des résultats à l'échelle de Wechsler d'intelligence (version WAIS) et des informations recueillies lors d'entretiens.

Le praticien dispose aussi du second logiciel, *SAMUEL*-Diagnostic, sur lequel sont enregistrées toutes les actions effectuées par le sujet.

➤ D'autres informations sur l'utilisation de SAMUEL

Plusieurs publications, postérieures à la publication du manuel, méritent d'être signalées car elles fournissent des indications concernant les utilisations possibles de ce logiciel *SAMUEL*. Un exemple d'utilisation de cette épreuve dans un cadre scolaire a été développé par l'auteur (Rozencwajg et Francequin, 1999). Cette publication correspond en grande partie au contenu des études de cas du manuel. *SAMUEL* a également été utilisé après d'adultes salariés (Rozencwajg *et al.*, 2005). Enfin, deux publications de Rozencwajg présentent, dans le cadre d'une approche intégrative de l'intelligence, les principaux éléments d'une évaluation reposant sur ce logiciel (Rozencwajg, 2005 et 2003).

Conclusion sur SAMUEL

La démarche utilisée par les concepteurs de ce logiciel *SAMUEL* nous semble très intéressante et elle constitue un bon exemple d'une application concrète de résultats de recherches à l'élaboration de tests d'aptitudes qui dépassent le simple constat d'un niveau de performance. Sans revenir sur les intérêts de *SAMUEL*, certains aspects de la démarche des auteurs mériteraient d'être précisés afin de mieux assurer la fiabilité de la catégorisation (des stratégies et des sujets) qui est effectuée :

1° La modélisation théorique des trois stratégies nous indique qu'il est relativement aisé de repérer la stratégie globale, mais que les stratégies analytique et synthétique, toutes les deux performantes dans cette situation, sont plus difficiles à distinguer sur les critères retenus. L'une des deux variables considérées comme pertinentes pour cette distinction (voir le tableau 6.1) est la fréquence des regards : un sujet « analytique » regarderait plus souvent le modèle (1 regard par cube en moyenne, donc une valeur théorique de 1 sur cet indice) tandis que le sujet « synthétique » regarderait moins souvent le modèle (1 regard par forme géométrique, soit un indice théorique de 0.20). On peut rappeler que, dans *SAMUEL*, le modèle n'apparaît qu'à la suite d'une demande du sujet (un clic de souris) et il disparaît dès que le sujet effectue une action (prendre un cube, par exemple). Rappelons qu'en situation classique de passation, le modèle est consultable à tout instant par le sujet. Les caractéristiques de ce dispositif informatisé peuvent alors avoir des effets sur cette fréquence des regards, comme par

exemple, inciter le sujet à mémoriser davantage le modèle[1] qu'il ne le ferait dans une passation classique. Une faible fréquence des regards ne pourrait donc plus être associée de manière systématique à une stratégie synthétique. D'ailleurs nous pouvons trouver dans le manuel des éléments d'observation qui viennent confirmer nos remarques :

> « Il faut souligner qu'un nombre non négligeable de personnes utilisant un ordre linéaire de placement des carrés conserve néanmoins l'ensemble du modèle en mémoire. » (Rozencwajg *et al.*, 1999, p. 25).

De plus, on peut également envisager que la signification d'un regard diffère en fonction des sujets : vérifier une action[2], préparer une action...

2° Dans le calcul du profil individuel, tous les indices ont le même poids (voir Rozencwajg *et al.*, 1999, p. 42), ce qui signifie que les auteurs considèrent que ces indices sont du même niveau de pertinence dans l'identification des stratégies. Pourtant nous pourrions envisager que certains indicateurs, et nous pensons ici plus particulièrement à l'ordre de placement des cubes, sont plus révélateurs que d'autres de la stratégie utilisée par le sujet. Il pourrait être pertinent d'envisager alors une pondération des indices dans le calcul du profil individuel.

3° Le sujet est catégorisé pour chaque item dans la stratégie la plus proche de son profil (en terme de distance). Mais que se passe-t-il quand ce profil est très éloigné des trois stratégies ? Autrement dit, le sujet est-il catégorisé quelles que soient les valeurs de ses indices ? *A priori*, d'après les informations du manuel, la réponse est positive. Une autre approche, qui nous semble plus satisfaisante, consisterait à définir une valeur minimale (un seuil) pour pouvoir catégoriser, avec plus de confiance, le sujet dans l'une des stratégies. De même, en cas de différences minimes entre deux profils stratégiques, il conviendrait de définir une valeur minimale de différence pour pouvoir, là encore, catégoriser le sujet. Une réflexion pourrait être menée à ce sujet.

4° Le diagnostic de SAMUEL ne repose que sur un nombre limité d'items, les quatre items à 9 carrés. Un nombre plus important d'items complexes permettrait, là encore, de mieux assurer la mesure.

1. De manière à lui éviter d'avoir à réaliser une nouvelle demande de consultation visuelle du modèle.
2. Un lien est possible ici avec des variables conatives : par exemple, on peut envisager qu'un sujet plus anxieux va vérifier plus souvent le modèle qu'un autre sujet utilisant pourtant la même stratégie...

Conclusion sur l'analyse des stratégies dans les tests

Comme nous venons de l'illustrer, l'identification des stratégies utilisées par un sujet est une problématique délicate, ce qui explique sans doute le faible nombre d'épreuves de ce type : mis à part *SAMUEL*, qui vient faire l'objet de cette présentation, il n'y a, à notre connaissance, aucun test disponible. On peut éventuellement signaler le test C.A.S[1] de Naglieri et Das, présenté par Kaufman (2001), qui vise également à identifier des stratégies mais, à l'inverse de *SAMUEL*, cet aspect ne semble pas essentiel dans les mesures effectuées par ce test. De plus le C.A.S n'est pas diffusé en France.

Pouvoir distinguer les sujets sur leur démarche de résolution, sur le type de stratégie qu'ils ont utilisé dans une épreuve, et donc au final sur leurs processus cognitifs, apporterait sans aucun doute aux praticiens des informations pertinentes, *qualitatives*, complémentaires aux constats classiques des tests qui ne reposent, le plus souvent, que sur des informations *quantitatives* relatives au niveau de performance. Ces informations peuvent être particulièrement utiles dans une perspective éducative ou de remédiation.

Cette approche cognitivo-différentielle, que l'on peut également qualifier d'évaluation diagnostique, centrée sur les processus cognitifs, et appliquée aux tests, mérite d'être développée car elle renseigne sur les processus mentaux en jeu dans ces épreuves. Elle constitue sans aucun doute l'une des perspectives les plus prometteuses de renouvellement des tests d'intelligence (voir également sur ce point Huteau et Lautrey, 1999a, chapitre VIII).

1. *Cognitive Assessment System.*

CHAPITRE
7

L'évaluation dynamique

Sommaire

APPARUE il y a une vingtaine d'années dans le paysage de la psychométrie, l'évaluation dynamique a été présentée (*cf.* Brown & French, 1979 ; Sternberg, 1985 ; Lidz, 1987) comme innovante et susceptible de renouveler les pratiques de diagnostic cognitif. De quoi s'agit-il et quelles méthodes existe-t-il ?

1. Les principes de l'évaluation dynamique

Définition

L'évaluation dynamique se distingue de l'évaluation conventionnelle, dite « statique », principalement sur deux aspects : la nature de ce qui est évalué et le rôle de l'examinateur.

Alors que l'évaluation classique mesure les capacités qu'un individu a développées, à un moment donné, et porte sur ses performances obtenues dans le test, l'évaluation dynamique vise à évaluer les capacités latentes, celles qui ne s'expriment pas spontanément, et cherche à apprécier la sensibilité du sujet à une situation d'apprentissage dans laquelle il est placé.

Il découle de cette différence d'objectif une différence de procédure de passation. Dans le cadre de l'évaluation classique, l'examinateur reste neutre, il ne doit pas influencer le sujet dans sa manière de répondre, ne pas donner de *feed-back* et établir de la façon la plus objective possible une mesure de ce que le sujet peut produire lorsqu'il est laissé à ses seules ressources. Au contraire, la caractéristique principale de l'évaluation « dynamique » est qu'elle combine évaluation et formation du sujet. Plusieurs procédures d'évaluation existent. Tous font intervenir l'examinateur pour fournir au sujet du *feed-back* et des aides lui permettant de progresser dans sa manière de résoudre les tests. Les progrès réalisés sont alors considérés comme des indicateurs de ce qui est appelé le « potentiel d'apprentissage » du sujet.

L'évaluation dynamique : les précurseurs

Deux auteurs ont particulièrement contribué à établir les bases théoriques de l'évaluation dynamique. Le premier est Vygotski dont les conceptions publiées en 1934 dans *Pensée et Langage* ont fortement influencé les

recherches dans le domaine durant ces vingt dernières années. Vygotski (1934/1985) introduit le concept de zone proximale de développement (ZPD) pour rendre compte de la marge d'éducabilité que possède l'enfant à un moment donné de son développement. Il s'agit de l'espace de développement possible, au-delà du niveau de développement actuellement atteint, que l'enfant pourra s'approprier avec l'aide d'un tiers. Le modèle de Vygotski met l'accent sur la médiation sociale du développement. La ZPD correspond donc à l'écart entre le niveau actuel de l'enfant, c'est-à-dire celui de sa performance autonome, et celui qu'il pourra atteindre s'il est aidé de façon adéquate.

Le second auteur dont la contribution à la définition de la notion d'évaluation dynamique a été particulièrement marquante est André Rey. La même année que Vygotski publiait *Pensée et Langage*, il publiait de son côté et de façon indépendante, un article intitulé « D'un procédé pour évaluer l'éducabilité » dans lequel il énonçait des principes très voisins du concept de ZPD en insistant sur la nécessité de « faire porter l'examen sur la forme et la vitesse des processus d'accommodation » (Rey, 1934, p. 299). Cette proposition s'accompagne d'une critique sévère des tests conventionnels qui évaluent des performances basées sur des apprentissages antérieurs dont on ne connaît rien. Cela conduit alors à comparer selon les mêmes critères des individus qui n'ont pas nécessairement bénéficié des mêmes occasions et conditions d'apprentissage, c'est-à-dire à comparer ce qui n'est pas comparable. Afin de rendre la comparaison possible, l'auteur proposait alors de placer les sujets dans des conditions standardisées d'apprentissage et de faire porter l'évaluation sur l'apprentissage lui-même, celui-ci rendant mieux compte de « l'éducabilité » de l'individu que ne le font les performances.

2. Les procédures d'évaluation

Le dispositif d'évaluation de « l'éducabilité » que proposait Rey était une tâche d'apprentissage d'un dispositif spatial découvert par tâtonnements (le test des plateaux) dans lequel la vitesse d'apprentissage était mesurée par le nombre d'essais nécessaires pour parvenir à la réussite. Par la suite, deux procédures d'évaluation se sont imposées (Büchel et Paour, 1990 ; Loarer et Chartier, 1996a) : la procédure Test-Apprentissage-Retest (T-A-R) et la procédure d'Aide au Cours du Test (ACT).

- *Dans la procédure T-A-R*, les performances des sujets sont initialement évaluées lors d'une première passation tout à fait classique. Vient ensuite une session d'apprentissage au cours de laquelle des explications sur la manière de résoudre les problèmes posés et sur la façon d'éviter les erreurs qu'ils ont commises lors du test sont fournies aux sujets. Une seconde passation de la même épreuve ou d'une version parallèle de la première épreuve est ensuite proposée. Le progrès du sujet entre le test et le retest donnera la mesure de son potentiel d'apprentissage.
- *Dans la procédure Aide au Cours du Test (ACT)*, la passation est unique et individuelle. Les aides sont fournies au sujet en cours de passation, à chaque item échoué. Le plus souvent ces aides sont standardisées et hiérarchisées, permettant à l'opérateur de donner d'abord des indices minimaux puis de les enrichir progressivement si nécessaire. La mesure du potentiel d'apprentissage tient alors compte de la quantité et de la nature des aides fournies et des réussites qui en découlent.

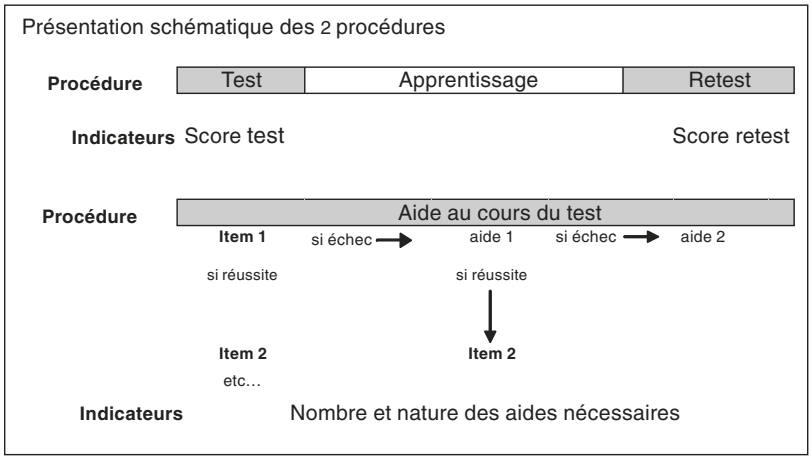

Figure 7.1
Présentation des deux procédures.

3. Les conceptions du potentiel d'apprentissage

L'idée de prendre en compte dans l'évaluation non seulement le niveau de performance de l'individu mais également la dynamique de son apprentissage

a séduit de nombreux auteurs qui y ont vu une manière d'améliorer la validité de la mesure de l'intelligence. Plusieurs revues de questions ont été consacrées à ce thème (Büchel et Paour, 1990 ; Büchel, 1995 ; Haywood & Tzuriel, 1992 ; Grigorenko et Sternberg, 1998 ; Laughton, 1990 ; Lidz, 1987). L'analyse de ces nombreux travaux fait émerger des conceptions du potentiel d'apprentissage qui divergent selon les auteurs, notamment pour ce qui est de ses rapports avec l'intelligence. Nous en avons principalement identifié trois :

- *La première* : le potentiel d'apprentissage serait un reflet plus pur de l'intelligence que celui fourni par les tests conventionnels mais n'en serait pas de nature très différente ;
- *La seconde* : le potentiel d'apprentissage recouvrirait strictement la notion de « Zone Proximale de Développement » de Vygotski et se démarquerait donc de l'intelligence évaluée par les tests statiques ;
- *La troisième* : défendue par Feuerstein qui décrit sous le terme de « modifiabilité cognitive », une entité à la fois distincte de la ZPD de Vygotski et de l'intelligence classiquement évaluée.

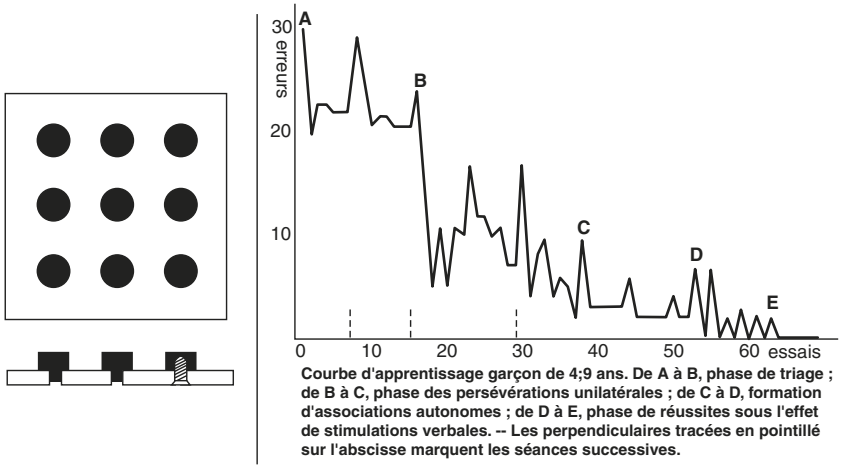

Figure 7.2
Les tests de plateaux d'André Rey.

Le test des plateaux d'André Rey (1934) : Le test est constitué de 4 « plateaux » : petites planches carrées de 14 cm de côtés. Chaque plateau est percé de 9 trous dans lesquels viennent s'encastrer des petits disques. Ces disques sont amovibles sauf l'une d'entre eux. L'apprentissage consiste à

apprendre à localiser ce dernier sans se tromper, alors que son emplacement est différent pour chaque plateau. Le temps nécessaire pour y parvenir et la courbe décroissante des erreurs sont des indicateurs de la capacité d'apprentissage.

Le potentiel d'apprentissage comme meilleure mesure de l'intelligence

Cette première conception est partagée par les auteurs qui, comme André Rey, estiment que les tests classiques ne disent rien sur les conditions dans lesquelles ont été réalisés les apprentissages antérieurs permettant de les réussir, et qui pensent que l'on peut obtenir une meilleure mesure du potentiel de l'individu en l'évaluant dans une situation d'apprentissage dont on contrôle les paramètres.

C'est la position défendue par Milton Budoff et ses collaborateurs (Budoff, 1987 ; Budoff et Corman, 1974 ; Budoff et Hamilton, 1976) dont la figure 7.3 illustre le modèle.

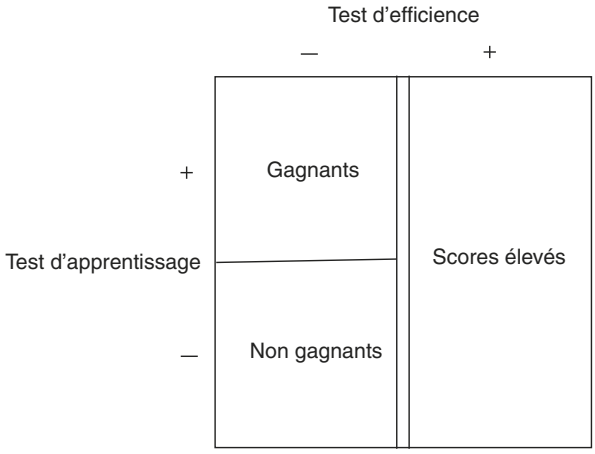

Figure 7.3
Modèle de Budoff (d'après Loarer, 2001).

L'objectif des recherches de Budoff était d'améliorer le diagnostic du retard mental en évaluant le potentiel d'apprentissage de sujets de faible niveau intellectuel (QI<90). Dans ses premiers travaux, il cherche ainsi à distinguer les sujets qui sont capables de tirer profit des aides qu'on leur apporte (les *gagnants),* de ceux qui n'y parviennent pas (les *non-gagnants*). Les capacités

des « gagnants » auraient été sous-évaluées par les tests conventionnels qui ne permettent pas de faire cette distinction entre « handicap éducatif » et « handicap réel ». Suite aux critiques faites notamment par Lidz (1991) concernant la délimitation des catégories proposées, il a, dans ses travaux les plus récents, substitué aux catégories des échelles d'évaluation continues, sans que cela n'affecte les principes de son modèle. Il utilise, pour l'évaluation dynamique, des adaptations de tests de facteur G et considère le potentiel d'apprentissage comme une mesure particulière du facteur G (Budoff, 1968). L'évaluation dynamique étant moins sujette aux biais socioculturels, il propose de la substituer à la mesure classique pour l'évaluation des enfants concernés par l'éducation spéciale.

Le potentiel d'apprentissage comme mesure de la zone proximale de développement

Vygotski (1934/1985, p. 269) introduit l'idée que :

> « Le psychologue doit nécessairement, pour déterminer l'état du développement, prendre en considération non seulement les fonctions venues à maturité mais aussi celles qui sont au stade de la maturation, non seulement le niveau présent mais aussi la zone proximale de développement. »

Il précise ensuite que le niveau présent de développement est évalué à l'aide de problèmes que l'enfant doit résoudre seul, et que la zone proximale de développement est déterminée par l'écart entre ce niveau et celui que peut atteindre l'enfant lorsqu'il est aidé par un adulte ou quelqu'un de plus compétent que lui. Les deux procédures d'évaluation (statique et dynamique) sont clairement présentées comme complémentaires. Elles donnent accès à deux facettes du développement cognitif : l'état du développement déjà réalisé grâce aux apprentissages antérieurs, mais aussi l'étendue de la zone dans laquelle les futurs apprentissages pourront donner lieu à de nouveaux développements. Or, affirme encore Vygotski :

> « La zone de proche développement a une signification plus directe pour la dynamique du développement intellectuel et la réussite de l'apprentissage que le niveau présent de leur développement. » (p. 270)

Ni Vygotski ni ses collègues n'ont véritablement apporté de validation expérimentale à cette affirmation (Grigorenko et Sternberg, 1998), mais le modèle proposé a inspiré de nombreux psychologues, non seulement dans les pays de l'Est, en Russie et en R.D.A. (Guthke, 1990, 2000 ;

Rubtsov, 1981) mais aussi aux États-Unis (Brown, & French, 1979 ; Campione & Brown, 1987 ; Rogoff & Wertsch, 1984 ; Wertsch et Tulviste, 1992). Pour ces auteurs, le potentiel d'apprentissage est différent de l'intelligence classiquement évaluée. Ainsi, par exemple, la conception de la complémentarité des mesures statiques et dynamiques apparaît clairement dans une recherche menée par Campione & Brown (1987). Les auteurs ont réalisé auprès d'enfants d'âge préscolaire une étude sur la validité prédictive d'un test de QI (le WPPSI) et d'un test de potentiel d'apprentissage (procédure T-A-R des Matrices de Raven). Intégrant dans une régression multiple comme critère le score de gain résiduel au test d'apprentissage et, comme prédicteurs, le nombre d'aides dans ce test et le score de QI au WPPSI, ils concluent que les deux scores ne se confondent pas, puisque chacun explique une fraction différente de la variance des scores de gain.

De même, Guthke et collaborateurs qui ont développé à Leipzig depuis le milieu des années soixante des travaux sur les « tests d'apprentissage », proposent de séparer (Guthke, 1980 cité par Guthke, 1992) ce qu'ils appellent l'*intellectual status* que l'on pourrait traduire par statut (ou état) intellectuel et ce qu'ils appellent le potentiel intellectuel *(intellectual potential).*

Le potentiel d'apprentissage comme évaluation de la modifiabilité cognitive

La position de Feuerstein est née de sa pratique. Il a été amené à évaluer, en Israël, des adolescents immigrants, issus de pays différents et pour la majorité peu scolarisés. Jugeant les tests conventionnels trop fortement marqués culturellement et inaptes à guider des interventions de remédiation, il a élaboré deux programmes complémentaires :

– *l'un* visant à l'évaluation du potentiel d'apprentissage, le *learning Potential Assessment Device* (LPAD ; Feuerstein, Rand, & Hoffman, 1979 ; Jensen, & Feuerstein, 1987) ;
– *l'autre* visant à la remédiation cognitive, le *Programme d'Enrichissement Instrumental* (PEI ; Feuerstein, Rand, Hoffman, & Miller, 1980).

Les deux démarches sont indissociables et visent toutes deux, grâce à la médiation mise en œuvre, à augmenter la *modifiabilité cognitive* du sujet. Le but du LPAD est de fournir, par une intervention courte, les éléments qui vont guider l'intervention plus lourde menée par le PEI.

Constitué d'une batterie de quinze épreuves verbales et non verbales inspirées de tests classiques, le LPAD structure une démarche clinique devant aboutir à un bilan qualitatif des fonctions cognitives et à des préconisations pédagogiques. Le but de l'application du LPAD – expliquent Feuerstein, Rand, Jensen, Kaniel, & Tzuriel (1987, p. 42) – est de provoquer des modifications cognitives structurales afin d'en établir les limites tant quantitatives que qualitatives.

Contrairement aux épreuves d'apprentissage proposées par les autres auteurs, la passation est peu standardisée. Le nombre d'épreuves, leur durée, la nature et la quantité de médiation fournie dépendent largement des caractéristiques du sujet et de la perception que l'examinateur en a. Le concept d'expérience d'apprentissage médiatisé (*Mediated learning experience* (MLE) ; Feuerstein *et al.*, 1980) est très voisin de celui de médiation sociale du développement proposé par Vygotski. Feuerstein explique les déficits cognitifs comme résultant d'un manque de MLE et envisage de remédier à ces déficits en apportant, à l'occasion de tâches proposées dans le LPAD et dans le PEI, une bonne médiation. Bien que s'inspirant très largement de la théorie de Vygotski, Feuerstein ne fait pas explicitement référence à la notion de ZPD et ne positionne pas l'évaluation du potentiel d'apprentissage comme complémentaire de l'évaluation statique. La définition de la modifiabilité cognitive comme une capacité très générale d'auto-adaptation et d'apprentissage évoque, il est vrai, de nombreuses définitions habituellement données de l'intelligence.

4. Les objectifs de l'évaluation dynamique

De l'ensemble des travaux menés ces dernières années et de ces différentes conceptions évoquées, nous identifions principalement quatre grands objectifs distincts, et parfois complémentaires, poursuivis par les auteurs qui se proposent de mener, à l'aide des dispositifs décrits, des évaluations dynamiques.

1r objectif : Améliorer la mesure de l'intelligence

L'évaluation dynamique est supposé permettre l'obtention d'une mesure plus complète et plus valide de l'intelligence que celle réalisée par

l'évaluation traditionnelle. Plusieurs arguments sont avancés pour étayer cette affirmation.

- Elle permet de limiter l'effet du hasard dans les réponses. L'évaluation donnant lieu généralement à une double mesure (procédure T-A-R), la probabilité qu'une bonne réponse soit obtenue au hasard est réduite ;
- Elle permet de corriger les biais socioculturels qui affectent la mesure classique. Les occasions d'apprendre ayant pu être différentes suivant les personnes, la phase d'apprentissage incluse dans le dispositif d'évaluation fournit à tous une égale opportunité de familiarisation aux tâches. Le résultat obtenu après entraînement refléterait donc mieux l'aptitude à raisonner que la performance spontanée produite sans aides ;
- Elle permet de distinguer entre vrai et pseudo-déficit. Les travaux de Budoff et col. (Budoff & Corman, 1974 ; Budoff et Hamilton, 1976 ; Budoff, 1987) illustrent cette propriété de la mesure du potentiel d'apprentissage. Ayant fait passer une épreuve de potentiel d'apprentissage à des sujets présentant des troubles de comportement ou ayant une vie familiale perturbée, les auteurs distinguent trois types de sujets : les « gagnants », se montrant capables de bénéficier d'un apprentissage, les « non-gagnants », qui ne profitent pas ou très peu d'un apprentissage, et les sujets à « scores élevés », ainsi nommés en raison de leur score élevé dès le prétest. Les sujets des deux premiers groupes ayant des scores faibles au prétest, n'auraient pas été distingués par une évaluation statique. Or, ils n'ont pas la même capacité à profiter de la situation d'apprentissage et l'observation de leurs comportements ultérieurs le confirme. Aussi, les auteurs concluent-ils que la mesure du potentiel d'apprentissage est une mesure de l'intelligence plus fiable que la mesure traditionnelle ;
- Elle est plus complète parce qu'elle intègre les fonctions intellectuelles en cours de développement (*cf.* Vygotsky) et permet de mesurer directement deux composantes de l'intelligence décrites comme importantes selon les théories cognitivistes du traitement de l'information : la vitesse d'apprentissage et l'efficience du transfert (Brown & Ferrara, 1985, p. 286).

2e objectif : Évaluer l'éducabilité cognitive de l'individu

La visée première de l'évaluation dynamique est, pour plusieurs auteurs, la mesure de l'éducabilité des individus. Le potentiel d'apprentissage est alors

pris comme prédicteur du développement cognitif ultérieur de la personne. Cette conception rejoint directement celle de Vygotsky (1935/1985) qui affirmait que la mesure du niveau de développement potentiel a une valeur prédictive plus grande pour la dynamique du développement intellectuel que la mesure du niveau actuel de développement.

Sous le terme d'éducabilité, Rey, pour sa part, considère l'adaptabilité du sujet à une situation nouvelle. Il se propose de la mesurer à travers la qualité et la vitesse de l'apprentissage d'une tâche d'exploration et de localisation spatiale.

De même, selon Feuerstein, le but de l'application du LPAD est la mesure de la « modifiabilité cognitive » des sujets qu'il décrit comme la possibilité que possède tout individu de « se modifier » et de former « de nouvelles structures cognitives » qui n'étaient pas auparavant dans son répertoire (Feuerstein, 1990, p. 123).

De façon concrète, Budoff (cité par Dias, 1991) utilise des tests de potentiel d'apprentissage pour intégrer des élèves de classes spécialisées dans des classes dites normales.

3e objectif : Pronostiquer la réussite dans les apprentissages ultérieurs

Les tests d'aptitudes classiques sont fréquemment utilisés en bilan d'orientation pour pronostiquer la réussite scolaire ou l'adaptation à des formations professionnelles. Ils remplissent d'ailleurs assez bien cette fonction. Dans ce cas, le pronostic des acquisitions futures est fait à l'aune du niveau des acquisitions antérieures, et donc du constat actuel. Pourtant, nombreux sont les auteurs qui critiquent cette démarche (*e.g.* Wagner & Sternberg, 1984) et certains considèrent plus valide pour diagnostiquer les capacités d'apprentissage d'une personne de la placer directement en situation réelle d'apprentissage. Le pronostic d'apprentissage n'est plus alors fondé sur un échantillon de performances mais sur un échantillon d'apprentissage. Cette recherche d'une meilleure homogénéité de contenu entre la variable observée et la variable prédite est de même nature que celle qui motivait les expériences menées dans les années vingt (décrites par Caroll, 1962, cité par Hurtig, 1995) où l'on faisait apprendre aux enfants des langues artificielles pour estimer leur aptitude à l'apprentissage des langues étrangères. La démarche est également dans l'esprit des tests *in basket* qui consistent à prélever un échantillon de la situation pour laquelle on cherche à prédire l'adaptation

du sujet et le proposer sous la forme d'un test standardisé. L'évaluation dynamique présenterait donc une meilleure validité de contenu vis-à-vis des apprentissages ultérieurs.

Une seconde raison menant à préférer l'évaluation dynamique pour le diagnostic de l'adaptation aux situations futures de formation est avancée par Budoff (1987). L'évaluation dynamique offre une meilleure conformité aux situations réelles de la vie, propriété que l'on pourrait qualifier de meilleure validité écologique. En effet, les conditions de standardisation des épreuves classiques placent le sujet dans un contexte artificiel ayant, à de nombreux égards, peu à voir avec les situations naturelles qu'il aura ensuite à affronter (voir Paour *et al.*, 1995, p. 64). On notera particulièrement l'absence de possibilité qu'a le sujet, dans la démarche classique, de tirer profit de ses erreurs au cours de la passation. Or, dans la vie courante la capacité à tenir compte du produit de ses propres actions est un facteur important d'apprentissage. L'évaluation dynamique lui permet de se manifester.

4e objectif : Recueillir des indications utiles à l'intervention pédagogique

La finalité pédagogique de l'évaluation dynamique est exprimée par de nombreux auteurs. Contrairement à l'évaluation classique de l'intelligence dont l'incapacité à fournir des indications utiles pour l'enseignant a de nombreuses fois été soulignée, il semble qu'il y ait une passerelle naturelle entre mesure du potentiel d'apprentissage et intervention pédagogique.

> « Les méthodes d'éducation cognitives constituent le prolongement naturel et nécessaire des instruments d'évaluation dynamique des aptitudes. » (Paour *et al.*, 1995, p. 86)

L'évaluation dynamique, affirme Das (1987) est relié de façon symbiotique aux concepts d'intervention et d'enrichissement. Cette finalité pédagogique de l'évaluation dynamique est également très explicitement affichée par Feuerstein (Feuerstein *et al.*, 1979, 1987). Alors que la plupart des auteurs (voir par exemple Guthke, 1990) séparent scrupuleusement le temps du diagnostic d'apprentissage du temps de l'intervention pédagogique, Feuerstein considère ces deux aspects comme indissociables et fait débuter l'intervention pédagogique au cours de la phase d'évaluation. L'intervention devient alors un moyen d'évaluation. Le but de l'application du LPAD est, dit-il (1987, p. 42) de provoquer des modifications cognitives structurales afin d'en établir les limites tant quantitatives que qualitatives. Il prévoit

en outre que l'évaluation se prolonge par un programme de remédiation cognitive (le PEI) permettant de remédier aux déficiences des fonctions cognitives constatées chez le sujet.

Les mérites de l'évaluation dynamique

On l'aura compris à la lecture de ce qui précède, l'évaluation dynamique est supposé présenter des avantages multiples sur l'évaluation classique.

La composante d'apprentissage introduite dans le test est considérée comme un moyen de détecter les possibilités d'évolution future des capacités de l'individu et, dans la mesure où elle neutralise les différences de familiarité avec la situation d'évaluation, de limiter les biais socioculturels. Ce serait donc une meilleure base de pronostic de la réussite et du développement ultérieurs que le test classique, notamment pour les sujets culturellement défavorisés ou issus d'une culture différente de celle dont les tests sont issus.

D'autres avantages sont également soulignés. L'évaluation dynamique permettrait, mieux que le test classique l'observation du fonctionnement du sujet (Paour *et al.,* 1995), ce qui correspond à une préoccupation de la psychologie cognitive contemporaine. Elle produirait également des informations plus directement utilisables dans une perspective éducative que le test classique (Campione & Brown, 1987). Elle serait enfin pour le sujet moins anxiogène et moins stressante (*cf.* Flammer & Schmid, 1995) et améliorerait son sentiment de compétence (Budoff, 1987).

Pourtant, si l'évaluation dynamique présente tous ces avantages sur l'évaluation classique, on peut se demander pourquoi elle ne l'a pas supplantée depuis longtemps et n'est pas aujourd'hui mieux inscrite dans les pratiques. La raison est, nous semble t-il à rechercher du côté des difficultés diverses que soulève ce type d'approche. Nous avons repéré, d'une part, des problèmes relatifs à la mesure du potentiel d'apprentissage :

- *Problèmes pratiques*, relatifs à la mise en œuvre de protocoles d'évaluations plus complexes et plus lourds que dans l'évaluation classique ;
- *Problèmes méthodologiques* liés principalement à la difficulté de mesurer le changement et à garantir les qualités métrologiques de l'évaluation dynamique ;
- *Problèmes théoriques* qui concernent la nature même de ce qui est évalué. Il est de toute évidence artificiel de dissocier radicalement les aspects méthodologiques et théoriques. Ainsi, par exemple, la façon dont les différents auteurs envisagent de prendre en compte les qualités

métrologiques de l'évaluation dépend de la conception qu'ils ont du potentiel d'apprentissage.

Nous examinerons ces différentes difficultés et présenterons quand ils existent des éléments de réponse, issus de recherches actuelles.

5. Les difficultés pratiques et méthodologiques de l'évaluation dynamique

Nous ne nous appesantirons pas sur les difficultés rencontrées par les praticiens confrontés à la mise en œuvre de procédures nettement plus lourdes que pour l'évaluation statique. Il est vrai que l'évaluation du potentiel d'apprentissage s'inscrit dans une certaine durée puisqu'il convient d'ajouter à la durée de l'évaluation la durée de l'apprentissage ou des aides (procédure ACT), ainsi que la durée du retest (procédure T-A-R). Elle peut également nécessiter une passation individuelle et des observations plus fines (ACT) et parfois une formation ou une expertise particulière. Nous suivons cependant volontiers Huteau et Lautrey (1999a) lorsqu'ils soutiennent que « les problèmes pratiques ne constituent pas un obstacle insurmontable au développement de l'évaluation dynamique » (p. 267). L'acceptation de ces contraintes pratiques est à notre avis strictement dépendante de la « valeur ajoutée » que peut apporter ce type d'évaluation et dépend donc de la résolution des problèmes méthodologiques et théoriques.

Les problèmes méthodologiques tiennent principalement au fait que la mesure du potentiel d'apprentissage est une mesure de changement. Elle présente de ce fait des risques de biais de mesure (Bacher, 1967 ; Embretson, 1987). Différents indices du potentiel d'apprentissage sont envisageables et permettent différents niveaux de contrôle des biais de mesure (Loarer et Chartier, 1994, 1996 ; Loarer, 2000 ; Huteau et Lautrey, 1999a).

Les problèmes rencontrés ne sont pas identiques selon la procédure utilisée : T-A-R ou ACT.

Problèmes méthodologiques relatifs à la procédure ACT

La procédure d'aide au cours du test suppose l'intervention du psychologue à des moments clés de la passation afin d'apporter au sujet une aide adaptée à la résolution d'une difficulté particulière sur laquelle il bute pour résoudre la tâche. Le choix du moment d'intervention et la nature de l'aide apportée

peuvent dépendre de l'appréciation du psychologue, ce qui rend alors la procédure peut standardisable et destine l'épreuve uniquement au cadre d'une intervention clinique. Dans ce cadre, on ne pourra attendre de l'épreuve qu'elle présente les qualités métrologiques classiquement attendues d'un test.

Certains auteurs ont souhaité standardiser la procédure d'introduction des aides ainsi que la nature même de ces aides. C'est le cas de l'épreuve de Ionescu présentée ci-dessous. Néanmoins, même dans ce cas, plusieurs problèmes méthodologiques demeurent. Nous citerons en particulier la difficulté à interpréter les indices de performance. Par exemple, le nombre de réussites consécutives à une aide, indice qui peut refléter la capacité du sujet à tirer profit de l'aide (et donc son « potentiel d'apprentissage »), est fortement dépendant du nombre d'aides fournies et donc du niveau initial de réussite aux items. Il faut échouer à l'item pour se voir proposer l'aide correspondante. Le potentiel d'apprentissage devient alors artificiellement corrélé négativement avec le niveau de réussite initiale. Le calcul d'un rapport « aide réussie/aide fournie » ne résout que très partiellement le problème. Par ailleurs, les aides étant fournies en cours d'épreuve, le score de réussite spontanée à un item inclut les effets des aides éventuellement données aux items précédents. Cette procédure ne permet donc pas de disposer d'une mesure très pure du niveau initial du sujet. Enfin, cette procédure ne peut que très difficilement aboutir à des mesures fidèles. C'est ce que démontrent de nombreuses études. Cette faiblesse de fidélité peut en particulier tenir au fait que les scores d'aides ne se distribuent souvent pas normalement, ce qui affecte le calcul de coefficients de fidélité.

Compte tenu de ces difficultés nous recommandons de réserver le recours à cette procédure à une approche clinique de l'évaluation, notamment lorsqu'il s'agit de détecter un potentiel individuel à apprendre, sans que l'on ait le souci d'une comparaison quelconque avec d'autres sujets ou de référence précise avec des critères externes, ou encore lorsque l'on s'intéresse principalement au rapport subjectif du sujet aux situations de résolution de problème et d'apprentissage.

Problèmes méthodologiques relatifs à la procédure T-A-R

La procédure T-A-R, évite certaines difficultés rencontrées dans la procédure ACT. Elle présente cependant également, comme nous allons le voir, différentes difficultés relatives au choix et à l'interprétation des indices, ainsi qu'en ce qui concerne la fidélité de ces indices.

➤ Les indices de potentiel d'apprentissage

Le potentiel d'apprentissage peut être mesuré par le gain (G) entre le test (X) et le retest (Y), donné par la différence Y–X. Cependant, ce gain présente une faible fidélité.

Pourquoi les scores de différences sont-ils peu fidèles ?

Comme nous l'avons vu dans le chapitre 2 de cet ouvrage, selon la théorie classique des tests, le score du sujet obtenu à un test (score observé) peut être décomposé en un score vrai et une erreur de mesure.

score observé = score vrai + erreur de mesure

Lorsque l'on procède à 2 mesures, on obtient deux scores observés (SO_1 ; SO_2) et chacun est composé d'un score vrai (SV_1 ; SV_2) et d'une erreur de mesure (E_1 ; E_2).

Lorsque l'on calcule la différence entre deux scores observés, les erreurs de mesures ne se soustraient pas mais se cumulent

$$\mathbf{SV_2 - SV_1 = (SV_2 - SV_1) + (E_2 + E_1)}$$

Le score de différence est donc affecté d'une variance d'erreur supérieure à celle de chacun des scores pris en compte.

La meilleure façon, dans l'absolu, de résoudre les problèmes liés à la mesure du changement est de faire appel aux modèles de réponse à l'item (*Item Response Theory*) appelés aussi « modèles à traits latents » (Dickes, Tournois, Flieller & Kop, 1994 ; Embretson 1987, 1989, 1991, 1995 ; Hambleton, Swaminathan & Rogers, 1991 ; Hambleton & Slater, 1997 ; Vrignaud, 1994, 1996). Ces modèles supposent l'existence d'un *continuum* latent sur lequel sujets et items peuvent être situés. Ils permettent de placer sur une échelle commune (le paramètre d'aptitude) les items du pré-test et du post-test et résolvent ainsi les effets de régression et les problèmes de fidélité. Ils permettent, en outre, d'estimer séparément le niveau de difficulté des items et le niveau de compétence des individus, ce qui est commode pour évaluer des progrès. Dans cette approche, on peut considérer le gain individuel du paramètre d'aptitude comme une mesure du potentiel d'apprentissage. Embretson (1991) a proposé un modèle multidimensionnel adapté à la mesure du potentiel d'apprentissage qui distingue deux variables unidimensionnelles : l'aptitude du sujet d'une part et sa modifiabilité d'autre part.

La mise en œuvre des modèles IRT est cependant délicate. Ils reposent sur des axiomes (d'unidimensionnalité, d'indépendance locale, etc.) qui

sont rarement satisfaits dans les situations concrètes et leur mise en œuvre nécessite un nombre élevé de sujets.

Il est donc utile d'envisager d'autres indices reflétant le potentiel d'apprentissage qui présenteraient moins d'inconvénients que le score de simple différence mais seraient plus opérationnels que ceux qui s'appuient sur les modèles IRT.

On peut, par exemple, corriger les effets de régression vers la moyenne en calculant un score de gain résiduel. Le score de gain résiduel (GR) est la part du score observé qui n'est pas attribuable à la régression du pré-test sur le post-test. La démarche consiste à calculer un score attendu Y' grâce à l'équation de la droite de régression des scores au retest sur les scores au pré-test, pour tous les sujets ayant un score donné au pré-test, et de calculer ensuite la différence entre ce score attendu Y' et le score observé Y_{obs}.Ce score ne permet cependant pas de distinguer entre deux sources de gains : celle qui est liée à la séance d'apprentissage (effet d'apprentissage) et qui peut concerner les principes logiques sollicités dans la tâche, et celle qui est liée à la répétition de la passation du test (effet du retest) et qui découle d'une meilleure familiarisation à la situation et du temps gagné par le sujet dans les items dont il se souvient. Or, on peut penser que ces deux effets n'ont pas le même sens ni la même capacité à prédire les apprentissages futurs.

Cela nous a amenés à proposer un nouvel indice (Loarer & Chartier, 1994) que nous avons appelé score de gain résiduel différentiel (GRD) qui consiste à estimer le score attendu Y' non plus sur le groupe expérimental, mais sur un groupe contrôle ne bénéficiant pas de la séance d'apprentissage. Le pronostic calculé par rapport à ce groupe (soit $Y'_{cont} = a'X+b'$) donne l'effet propre du retest. Pour un sujet du groupe expérimental, le score de potentiel d'apprentissage sera la différence entre le score attendu s'il avait fait partie du groupe contrôle Y'_{cont} et le score observé $Y_{obs.}$ Il s'agit d'un gain hypothétique, représentant la part de la note observée non attribuable à l'effet de retest. L'avantage de cette mesure est donc d'isoler l'effet de la séance d'apprentissage. L'inconvénient est la lourdeur du dispositif d'évaluation qui la destine principalement à la recherche.

Le score au retest apparaît cependant comme un compromis intéressant puisqu'il permet d'éviter les problèmes liés à la répétition de la mesure tout en étant d'obtention aisée. Il a néanmoins l'inconvénient de mêler dans un score global le niveau initial et le gain dû à l'apprentissage. Huteau et Lautrey (1999a, p. 256) proposent une façon élégante de séparer ces deux éléments lorsque l'on possède un critère extérieur, en calculant la corrélation partielle du post-test avec le critère lorsque la corrélation avec

le pré-test est partialisée, ou encore en réalisant une analyse de régression dans laquelle on introduirait successivement comme prédicteurs le pré-test puis le post-test. La fraction de variance supplémentaire expliquée par le post-test correspondant à l'effet propre de l'apprentissage. Cependant, cette pratique est réservée à des recherches et peu adapté aux pratiques classiques d'évaluation. En outre, dans de nombreuses études, la mise en œuvre de ce traitement est gênée par la présence de colinéarité entre les variables. À l'issue d'un ensemble d'études menées pour comparer les propriétés des différents indices de potentiel d'apprentissage, Loarer (2000) conclut que l'indice le plus commode à utiliser et le plus valide est bien le score au post-test.

La fidélité des mesures d'apprentissage

La fidélité des mesures d'apprentissage est menacée par plusieurs types de phénomènes :

1. *Les effets de plafonnement des scores* : Les épreuves de potentiel d'apprentissage sont fréquemment confrontées à des problèmes techniques liés à un « effet de plafond » : la marge de progression possible dans une épreuve n'étant pas infinie, les scores d'apprentissages peuvent s'en trouvent affectés. Par exemple, Büchel et al. (1990) cherchant à évaluer la stabilité, dans le temps, des gains entre test et retest se heurtent à un effet de plafond dans les apprentissages ;

2. *Les effets des erreurs de mesure* : Comme dans l'évaluation conventionnelle, la fidélité test-retest des épreuves d'apprentissage est relative aux erreurs de mesure aléatoires affectant l'observation (l'aptitude du sujet donnée par la "mesure vraie"). Lorsque le score d'apprentissage s'appuie sur deux scores (test et retest), les erreurs de mesure sont alors cumulées ;

3. *La stabilité du changement* : Dans l'évaluation dynamique, la fidélité est également dépendante de la stabilité du phénomène observé. La fidélité de la mesure du potentiel d'apprentissage suppose une stabilité dans la façon de changer, ce qui n'est pas toujours le cas.

6. Les problèmes théoriques de l'évaluation dynamique : que mesure t-on exactement ?

Au-delà des problèmes pratiques et méthodologiques qui ont été présentés et pour lesquels, nous l'avons vu, des réponses satisfaisantes semblent pouvoir être trouvées, un certain nombre de problèmes d'ordre théorique subsistent aujourd'hui et divisent les auteurs. Dans la période récente, de nombreux travaux portent sur l'évaluation dynamique et proposent des techniques de mesure du « potentiel d'apprentissage ». Cependant, le concept n'est pas toujours clairement défini et lorsqu'il l'est, les conceptions qu'en ont les différents auteurs diffèrent assez largement.

> « Préoccupés surtout par la construction d'instruments destinés à mesurer le potentiel d'apprentissage – expliquaient déjà Ionescu & Jourdan-Ionescu (1984, p. 920) – les chercheurs ont négligé les élaborations théoriques. »

Ce manque d'unité de vue sur la notion de potentiel d'apprentissage, encore présent aujourd'hui, donne parfois l'impression que l'on ne sait pas très bien ce que l'on mesure même si l'on sait parfaitement le mesurer.

Les rapports entre le potentiel d'apprentissage et l'intelligence

Certains auteurs ne voient pas la nécessité théorique de distinguer les deux dimensions. Pour eux, les tests classiques et les tests de potentiel d'apprentissage mesureraient, sous des formes différentes, la même chose. En effet, les tests classiques d'intelligence mesurant, à travers le niveau d'efficience actuel, le produit des apprentissages antérieurs, ils prendraient indirectement en compte le potentiel d'apprentissage. La mesure statique de l'intelligence intégrerait donc celle du potentiel d'apprentissage.

Lautrey (1994) fait remarquer que cette position ne devrait cependant pas nécessairement exclure l'intérêt d'une évaluation dynamique. En effet, compte tenu du fait que les occasions d'apprendre peuvent avoir été différentes suivant les individus, peut-être obtiendrait-on une meilleure évaluation de l'intelligence par une mesure directe de la capacité d'apprentissage qu'à travers ses produits. Nous noterons que dans ce cas, le recours à l'évaluation dynamique n'est alors envisagé que comme une possibilité que se donne le psychologue de compenser, pour mieux mesurer l'intelligence, certains biais culturels.

Pour d'autres auteurs, il semble que les deux dimensions soient fondamentalement distinctes. Les tests classiques et les tests de potentiel d'apprentissage mesureraient des réalités différentes. Pour Vygotsky, par exemple, et donc pour les auteurs qui s'en inspirent (Brown & Ferrara, 1985 ; Campione & Brown, 1987 ; Day, 1983), la zone proximale de développement débutant là ou finit la zone de développement actuel, les tests classiques et les tests de potentiel d'apprentissage mesureraient donc, par définition des entités psychologiques distinctes. Cette hypothèse semble être confirmée par les résultats obtenus par plusieurs auteurs (Guthke, 1982 ; Lidz, 1987) qui montrent que les scores de réussite spontanée (ou prétests) et les scores d'apprentissage (post-tests ou scores de gains) sont faiblement intercorrélés. Cependant Flammer & Schmid (1995, p. 193) expliquent que ces résultats peuvent être dus à des artefacts méthodologiques.

La nature et la signification des progrès consécutifs à l'apprentissage évalué

« Les fondements de l'évaluation dynamique s'appuient sur le postulat de l'éducabilité de l'intelligence » écrivent Paour *et al.* (1995, p. 47). Nous pouvons ajouter que si l'évaluation dynamique et l'éducation cognitive partagent les mêmes racines épistémologiques ils partagent également les mêmes ambiguïtés théoriques. Aussi, la question cruciale de la nature des effets induits par le programme d'éducation cognitive est posée ici à propos des progrès mesurés par le potentiel d'apprentissage. Les progrès « résultent-ils d'une transformation du sujet, d'une modification de sa représentation de la tâche et/ou d'une réduction de la complexité initiale de la tâche ? » s'interrogent très justement Paour *et al.* (1995, p. 82).

On a vu l'importance de cette question à propos de l'évaluation des effets de méthodes de remédiation cognitive (*cf.* Loarer, 1998). Faut-il interpréter les effets observés à l'issue d'un programme d'éducation cognitive consistant à entraîner les sujets à résoudre des problèmes extraits directement ou inspirés de tests d'intelligence, comme des indices de développement cognitif ou bien comme le simple résultat d'une familiarisation aux situations de tests ?

Tout comme pour les effets des méthodes d'éducation cognitive, la réponse à cette question passe par l'étude de l'intégration fonctionnelle de ce qui a été acquis à propos de certaines tâches (transférabilité proche et éloignée, immédiate et différée à des tâches différentes requérant une activité cognitive de même type) (*cf.* Huteau et Loarer, 1992). Seule une

telle étude permettra de dire si le sujet a seulement été entraîné à réussir au test ou a fait l'apprentissage de procédures cognitives nouvelles réutilisables ultérieurement et transférables à des situations analogues.

La nature des contenus et opérations cognitives qui sont évalués et entraînés dans le cadre de l'évaluation dynamique

Le reproche majeur fait aux tests d'intelligence est qu'ils ne permettent généralement pas de comprendre le fonctionnement cognitif des individus (voir chapitre 7). De ce fait, ils sont d'un faible secours dans l'établissement d'un diagnostic sur la nature des difficultés de fonctionnement ni dans la prescription de mesures de remédiation. À l'inverse, « l'évaluation dynamique s'inscrit dans cette démarche d'élucidation des contenus cognitifs des tests d'intelligence » (Paour *et al.*, 1995, p. 52).

La construction d'un test d'évaluation de potentiel d'apprentissage – et particulièrement l'élaboration des aides spécifiques fournies au sujet, ainsi que le choix de tâches de transfert – demande en effet une connaissance des domaines cognitifs à évaluer et une analyse fine des items proposés.

Quels que soient les objectifs poursuivis, l'évaluation dynamique ne peut se dispenser d'une analyse des contenus et des processus mobilisés dans la résolution des tâches proposées au sujet. Ceci est particulièrement vrai lorsque la finalité de l'évaluation est l'intervention psychopédagogique. Différentes démarches d'analyse des tâches cognitives ont été proposées (*cf.* Sternberg, 1977 ; Glaser & Pellegrino, 1978/79 ; Pellegrino, 1985), mais la complexité de leur mise en œuvre les rend essentiellement utilisables dans un contexte de recherche.

Sur ce point, l'évaluation dynamique rencontre des limites qui sont celles de l'avancement des recherches en psychologie cognitive. Bien entendu, le besoin de connaissances de ce type pour les applications psychométriques et pédagogiques peut constituer une incitation importante à ce que s'intensifient les travaux dans le domaine. Mais on sait aussi que ces recherches sont coûteuses et que la production de connaissances nouvelles est lente.

Nous noterons également que même si l'analyse des tests peut aboutir à la compréhension des processus cognitifs de résolution impliqués dans les tâches, elle ne dit pas comment on peut ensuite aider les individus à acquérir la maîtrise de ces processus lorsque l'on constate qu'ils leur font défaut. Sur

ce point, on se reportera aux travaux sur les méthodes d'éducation cognitive (voir par exemple Loarer *et al.*, 1995 ; Loarer, 2001).

La validation du potentiel d'apprentissage et des critères de validité

La notion de validité prédictive des épreuves de potentiel d'apprentissage est fondée sur le postulat que la capacité à apprendre qui se manifeste aujourd'hui dans une situation de test, se manifestera à nouveau demain dans des situations réelles. Ce postulat soulève à notre avis deux problèmes majeurs. Le premier concerne la généralité de la mesure réalisée. Le second sa fidélité.

➤ Généralité de la mesure

Au cours de l'évaluation dynamique, l'entraînement est réalisé dans un domaine donné et dans des conditions données. L'hypothèse que cet échantillon particulier d'apprentissage soit représentatif de tous les apprentissages que la personne sera amenée à effectuer nous paraît très audacieuse et à dire vrai peu fondée. Bien sûr, on constatera que les tâches retenues par la majorité des auteurs (Feuerstein, Guthke, Budoff, Ionescu, ...) pour servir à la fois de support à l'évaluation et à l'apprentissage sont des tests fortement saturés en facteur g. Est-ce à dire que l'aptitude à apprendre, avec l'aide d'un psychologue, à résoudre des tests de facteur G témoigne d'une capacité générale d'apprentissage ?

Les travaux menés en psychologie cognitive durant ces trente dernières années sont nombreux à conclure à l'importance des contenus et des contextes spécifiques dans l'acquisition et la mise en œuvre des procédures cognitives (Chi, 1978 ; Borkowski & Cavanaugh, 1979 ; Lautrey *et al.* 1986 ; Pignault, 2007). Ces résultats concernent directement la problématique de l'évaluation dynamique. On a vu également à plusieurs reprises (Loarer *et al.*, 1995 ; Loarer, 2001) comment ils justifiaient la révision des postulats de base de certaines méthodes de remédiation cognitive.

Une étude que nous avons menée afin de tester le degré de généralité ou de spécificité de la capacité à apprendre (Loarer & Chartier, 1994) renforce ce point de vue. Nous avons bâti, selon la procédure T-A-R trois épreuves de potentiel d'apprentissage, chacune explorant un domaine cognitif différent : raisonnement inductif, raisonnement spatial et créativité.

Appliquées aux mêmes sujets (123 adolescents de lycée professionnel et jeunes adultes en formation) les épreuves aboutissent à des mesures de la capacité d'apprentissage très faiblement corrélées entre elles, ce qui atteste de la spécificité du potentiel d'apprentissage. Les scores de potentiel d'apprentissage ne renvoient donc pas à une capacité générale à apprendre, mais à une capacité qui varie selon les domaines sur lesquels ont porté les apprentissages. Ce point de vue corrobore celui énoncé par Campione & Brown (1987), Brown & Ferrara (1987), ou encore Klauer (1990). Il semble donc nécessaire, ce qui limite sérieusement la portée de certaines épreuves, que pour réaliser une mesure de potentiel d'apprentissage à des fins pronostiques, le psychologue sélectionne soigneusement les tâches en relation avec les domaines visés.

La concordance entre l'épreuve de potentiel d'apprentissage et l'apprentissage ultérieur n'est pas seulement à considérer du point de vue des contenus. Elle est aussi à envisager sous l'angle du format de l'apprentissage. Ainsi, pour ce qui est de la prédiction de la réussite scolaire, la validité des tests d'apprentissage va dépendre de la concordance entre le mode d'entraînement et le mode d'enseignement. On ne voit en effet pas bien pour quelle raison le score de potentiel d'apprentissage obtenu après une séance d'entraînement très individualisé et donnant lieu à une forte médiation de la part du formateur constituerait un bon prédicteur de la réussite du sujet dans une formation ultérieure collective et faiblement médiatisée. C'est ce qui fait dire à de nombreux auteurs (Laughton, 1990 ; Jensen & Feuerstein, 1987 ; Lidz & Thomas, 1987 ; Flammer & Schmid, 1995) que les résultats scolaires ne sont pas toujours de bons critères de validation des scores de potentiel d'apprentissage. De même, pour ce qui est de l'éducabilité cognitive de la personne, nombreux sont les auteurs (par exemple Feuerstein *et al.*, 1979, 1998 ; Dias, 1991) qui soulignent la nécessité, pour que l'évaluation dynamique ait un sens, que la personne puisse continuer ultérieurement à se trouver dans un environnement favorable à son développement.

> « La perspective ouverte par la théorie du potentiel d'apprentissage ne peut être pleinement satisfaite que dans un environnement qui offre à ses membres les conditions de se modifier. » (Dias, 1991, p. 132).

Faute d'un tel environnement, la mesure du potentiel d'apprentissage restera la mesure non pas d'un développement futur mais... d'un potentiel futur non réalisé et de ce fait non validable.

Si l'on considère d'une part que les conditions habituelles de formation scolaire ou professionnelle sont rarement de même type que celles préconisées

par les auteurs pratiquant l'évaluation dynamique et d'autre part que l'évaluation du potentiel d'apprentissage est, comme nous l'avons dit, essentiellement pertinente pour les personnes qui vivent dans un environnement socioculturel défavorable, alors on aboutit à un certain paradoxe de la méthode : le potentiel d'apprentissage serait un bon prédicteur pour des situations dans lesquelles le sujet a peu de chances de se trouver placé. Ceci limite à notre avis l'étendue du domaine de validité prédictive du potentiel d'apprentissage.

➤ Fidélité de la mesure

Le pronostic d'apprentissage suppose une certaine stabilité dans la façon de changer et renvoie au problème de la fidélité de la mesure du potentiel d'apprentissage.

> « Si l'évaluation du potentiel d'apprentissage n'était pas fidèle, au moins à un moment donné, elle ne pourrait fonder la moindre activité diagnostique ou pronostique et elle serait donc strictement inutile » précise Lautrey (1994, p. 138).

Pourtant, cette propriété de la mesure a rarement été étudiée pour les épreuves de potentiel d'apprentissage, probablement en raison des problèmes techniques qu'elle pose et que nous avons évoqués.

À notre avis, cependant, le problème de fidélité de la mesure des potentiels d'apprentissages est plus théorique que méthodologique, notamment lorsqu'il s'agit de pronostiquer le développement (diagnostic d'éducabilité). Le pronostic de développement ultérieur fait implicitement référence à un modèle linéaire du développement cognitif. Lorsque Vygotski affirme que la ZPD renseigne mieux que le niveau actuel sur les apprentissages ultérieurs, il suppose une certaine stabilité individuelle des caractéristiques de la ZPD. Lorsque les auteurs contemporains suggèrent d'évaluer le potentiel d'apprentissage des individus, ils supposent également que celui-ci peut être considéré comme un trait caractéristique du sujet. La liaison recherchée est généralement étudiée par une régression statistique, simple ou multiple, de type linéaire. Or, aucune théorie génétique n'envisage aujourd'hui le développement cognitif comme un processus monotone. Nous pouvons même aller plus loin et souligner combien cette référence implicite est en contradiction avec certains postulats de base de l'éducabilité cognitive.

Dans le cadre de l'évaluation classique de l'intelligence, la validité de la prédiction s'appuie sur la stabilité dans le temps des caractéristiques

individuelles (relativement au groupe de référence). C'est le cas, par exemple du QI. De ce fait, le niveau futur peut être pronostiqué à partir du niveau actuel. Les tenants de l'évaluation dynamique postulent, au contraire, que la capacité d'apprentissage, la modifiabilité ou l'éducabilité n'est pas, pour un individu, une quantité fixée génétiquement une fois pour toutes, mais est susceptible de variations importantes en raison de multiples facteurs externes ou internes. Ainsi, par exemple, Feuerstein prétend provoquer par son intervention (LPAD ou/et PEI) une augmentation de la modifiabilité cognitive des individus et cela quel que soit leur âge (Feuerstein, 1980, 1990). De leur côté, Campione & Brown (1987, p. 87) soulignent la nécessité de réactualiser fréquemment la mesure du potentiel d'apprentissage. La mesure de l'éducabilité d'un individu, disent-ils, n'est valable que pour de brèves périodes parce qu'elle peut changer avec l'entraînement ou l'instruction. Feuerstein va plus loin et rejette l'idée même de fidélité dans l'évaluation du potentiel d'apprentissage au nom de l'instabilité du phénomène observé (Feuerstein *et al.*, 1987). Nous ne le rejoignons pas sur ce point car il devient alors inutile de tenter toute mesure.

Ce point de vue très optimiste ne prend pas en compte ce que les théoriciens du développement appellent les contraintes ou les limites développementales présentes dans toutes les théories du développement. Ce manque d'intégration de l'évaluation dynamique dans un modèle explicite du développement cognitif peut surprendre. De Ribaupierre (1995) explique ce phénomène par le clivage historique entre théories de l'apprentissage et théories développementales. Certains promoteurs de l'évaluation dynamique étant essentiellement des théoriciens de l'apprentissage, ils ont eu tendance à développer leurs conceptions en marge des grandes théories du développement et de ce fait à sous-estimer les contraintes structurales s'exerçant sur l'ampleur des progrès possibles.

7. Quels usages des épreuves de potentiel d'apprentissage ?

Quel que soit le modèle défendu, l'évaluation dynamique est supposé fournir les bases d'un meilleur pronostic des apprentissages ultérieurs que ne le fait l'évaluation statique. Aussi, les études comparant les validités prédictives, du point de vue de la réussite scolaire, de tests de potentiel d'apprentissage et de tests conventionnels de QI, sont assez nombreuses dans la littérature

(Grigorenko & Sternberg, 1998). Force est de constater qu'elles ne vont pas toutes dans le sens de l'hypothèse. Ainsi, par exemple, Sewell (1979, 1987) observe dans une étude de ce type menée en première année de primaire que la meilleure prédiction est donnée, pour l'ensemble de l'échantillon testé, par les tests conventionnels. Taylor & Richards (1990) arrivent aux mêmes conclusions : le Wisc-R s'avère être un meilleur prédicteur de la réussite scolaire en primaire que les tests d'apprentissage qu'ils ont utilisés. Une étude conduite par Guthke (1990) fournit également des résultats allant dans le même sens. Il constate, sur un échantillon de 400 enfants faisant l'objet d'un suivi durant leur scolarité primaire, que les résultats obtenus par un test classique de facteur G (MPC) prédit mieux la réussite scolaire (évaluée par les notes, les appréciations des maîtres et des tests de rendement scolaire) que ne le font les résultats d'un test de potentiel d'apprentissage (le RKL). Ce type de résultats a amené certains auteurs (par exemple Flammer, 1974, cité par Flammer & Schmid, 1982/1995, p. 204) à conclure qu'avec des sujets « normaux », l'apport de l'évaluation dynamique n'était pas suffisant pour justifier son coût supplémentaire.

Il n'en va cependant pas de même lorsque l'on s'intéresse aux sujets les plus faibles. Dans une étude de 1979, Sewell constate ainsi que, si le score classique de QI prédit mieux la réussite scolaire d'enfants blancs de classe sociale moyenne, c'est le score de retest de l'épreuve de potentiel d'apprentissage que fournit la meilleure prédiction pour un groupe d'élèves noirs de classe sociale défavorisée. De même, Guthke (1990) rapporte que lorsqu'il observe non plus l'ensemble de l'échantillon, mais seulement les élèves (5 %) qui avaient été signalés par la maîtresse de maternelle, à leur entrée en primaire, comme présentant un risque d'échec, c'est le score de potentiel d'apprentissage qui prédit le mieux leur réussite scolaire.

Les résultats que nous avons nous-mêmes obtenus dans l'épreuve des SPM, et présentés ci-dessous, vont dans le même sens. La séance d'aide ou d'apprentissage apparaît augmenter la validité de la mesure (score au retest). L'évaluation dynamique permet ainsi d'améliorer sensiblement le pronostic de réussite pour les sujets les plus faibles, alors qu'elle n'apporte aucune information supplémentaire concernant les sujets à niveau initial élevé. Ces deux éléments étayent, parmi les trois conceptions du potentiel d'apprentissage que nous avons décrites, celle qui voit dans la mesure du potentiel d'apprentissage une *amélioration de la mesure de l'intelligence*, notamment en limitant les biais socioculturels.

Ces résultats illustrent un paradoxe et s'accordent avec le point de vue de Budoff (1987) : les tests d'intelligence ont souvent été construits pour

repérer les déficits intellectuels et sont massivement utilisés pour l'évaluation des enfants déficients, c'est-à-dire pour l'usage dans lequel ils semblent les moins valides. Cela conforte la position de Budoff sur le sens à donner à la notion de potentiel d'apprentissage. Il apparaît clairement ici que l'évaluation dynamique présente un intérêt pour la compensation des biais socioculturels dans l'évaluation de l'intelligence.

L'un des avantages souvent cité par les défenseurs de l'évaluation dynamique est qu'elle permet de *recueillir des indications utiles à l'intervention pédagogique*. Nous disposons en l'état actuel de peu d'éléments probants allant dans ce sens. Il nous semble que des avancées ne pourront être faites dans cette voie :

– sans un effort de conception de nouvelles tâches d'évaluation permettant une analyse fine des stratégies mises en œuvre par les sujets (nous avons évoqué les limites des tâches adaptées de tests classiques pour analyser les erreurs des sujets) ;
– sans une avancée conjointe des connaissances sur les interactions entre individus et situation pédagogiques ;
– sans l'élaboration de situations de formation capables de fournir des critères fiables de validation des stratégies d'apprentissage repérées dans la situation de test.

8. Présentation d'épreuves

Les épreuves d'évaluation du potentiel d'apprentissage sont, à quelques exceptions près, peu diffusées et accessibles en France. Beaucoup ont été développées à l'occasion de recherches. Nous avons fait le choix ici de présenter trois épreuves :

- Une *première* épreuve adaptée des cubes de Kohs (Ionescu *et al.*, 1985, 1987, Loarer et Chartier, 1994) qui adopte la procédure ACT,
- Une *seconde* épreuve adaptée des Matrices de Raven (Loarer et Chartier, 1994, Loarer, 2001) qui utilise la procédure T-A-R,
- Une *troisième* épreuve adaptée d'un test de Faverge : le TEDE6 de Pasquier (2003) qui a opté pour une variante de la procédure T-A-R dans laquelle ne subsistent que les phases d'apprentissage de retest.

Seule cette dernière épreuve est disponible chez un éditeur. Les deux premières épreuves sont présentées ici afin de fournir des exemples prototypiques de matériels et de procédures d'évaluation dynamique.

L'épreuve de type « Aide au cours du test » de Ionescu et collaborateurs fondée sur les cubes de de Kohs

Une procédure d'évaluation dynamique basée sur l'épreuve des cubes de Kohs a été élaborée par Ionescu et al. (1985), Ionescu, Jourdan-Ionescu, & Alain (1987) et reprise et complétée par Chartier & Loarer (1994).

➤ L'épreuve

Le matériel utilisé a été construit à partir des neuf planches de l'épreuve de cubes de l'Échelle d'Intelligence de Wechsler pour adultes (WAIS-R).

Les principales caractéristiques de la procédure sont les suivantes :

- La passation est individuelle ;
- Chaque personne passe l'ensemble de l'épreuve, composée de 9 items ;
- Les aides ne sont données qu'en cas d'échec mais le sont jusqu'à l'obtention de la réussite ;
- Les aides sont standardisées et hiérarchisés, c'est-à-dire que l'opérateur commence par donner des indices minimaux, qui sont progressivement enrichis en cas d'échec ;
- Les aides sont fournies au sujet au cours de la passation en fonction des erreurs qu'il commet ;
- Une série de trois aides hiérarchisées est prévue pour chaque item (voir figure 7.4) :

  1. La première de ces aides consiste à présenter le modèle à l'échelle 1 (le modèle original est à l'échelle 1/2). Elle permet de compenser d'éventuels problèmes perceptifs ou des difficultés liés au changement d'échelle ;

  2. La seconde aide présente un modèle où sont tracées les limites des différents cubes, induisant une stratégie d'analyse de la figure en éléments séparés ;

  3. La troisième aide est une démonstration réalisée par le psychologue à l'aide des cubes.

En cas de réussite avec aide on revient systématiquement au modèle initial de l'item afin d'évaluer ce que les auteurs appellent le transfert d'apprentissage. Ce dernier constitue un aspect essentiel de la mesure du potentiel d'apprentissage. Il se réfère à la capacité qu'a la personne qui passe le test de profiter de l'aide, ou des aides apportées, non seulement pour réussir le niveau de tâche pour lequel l'aide a été apportée mais également d'exploiter le principe appris pour mieux réussir l'item de niveau de difficulté supérieur.

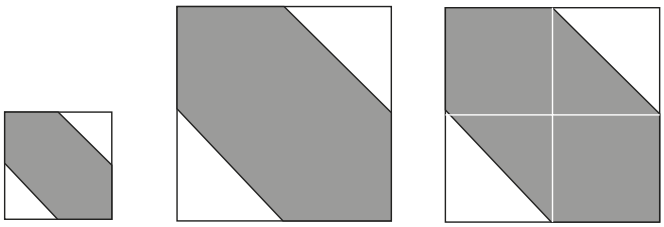

Figure 7.4
Principe des aides.

Les aides ne sont données qu'en cas d'échec, selon le schéma de passation décrit dans la figure 7.5**.** Quel que soit l'item considéré, le temps de réflexion du sujet est limité à 2 minutes pour la planche standard et à 1 minute pour chacune des aides.

➤ Les indices

Dans les études réalisées par Ionescu et al. auprès de déficients mentaux, trois notes ont été prises en compte, calculées soit à partir des réussites spontanées des sujets (NS, note spontanée) soit à partir du nombre d'aides efficaces fournies consécutivement à un item initialement échoué (NA, note d'aide) ou encore du nombre de réussites du modèle initial après aide (NT, note de transfert). Les auteurs considèrent la note spontanée comme équivalente à une mesure classique de l'aptitude. Cela n'est pas à notre avis tout à fait justifié, car l'effet d'apprentissage tient alors aussi bien à la familiarisation avec l'épreuve qu'aux aides éventuellement fournies. Quoi qu'il en soit, dans ces conditions, la prise en compte de la note de transfert dans un score global (NG = NS+NT) améliore quelque peu la validité prédictive de l'épreuve par rapport à un critère externe qui est la notation des moniteurs ayant eu à superviser le travail des sujets. Cette note globale explique 29 % de la

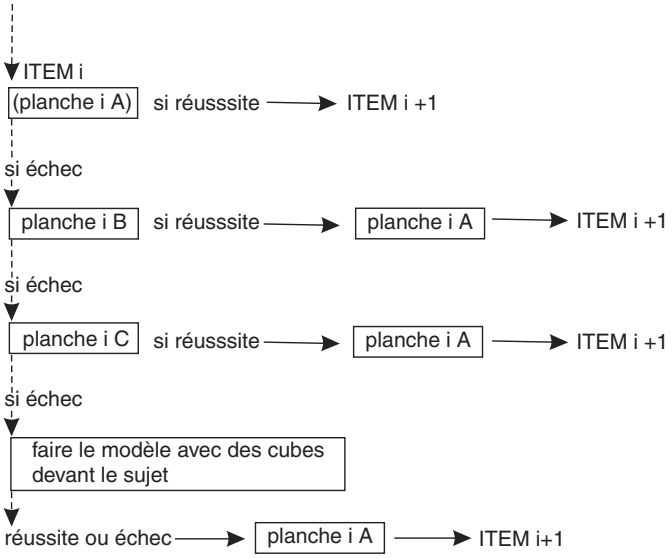

Figure 7.5
Schéma de la passation.

variance de l'Échelle de Compétence Professionnelle sur laquelle les sujets sont évalués.

Chartier et Loarer (1994) introduisent une variante en proposant 2 nouveaux scores par rapport à Ionescu et al. (1987) pour limiter le problème concernant la liaison négative observée habituellement entre le score de réussite spontanée et le nombre d'aides ou de transferts réussis. Il s'agit de deux rapports :

- *Le rapport (appelé RA, « rapport d'aide »)* donné par l'opération : nombre de réussites consécutives à une aide/nombre d'aides fournies, qui désigne la capacité du sujet à tirer parti des aides qui lui sont fournies ;
- *Le rapport (appelé RT « rapport de transfert »)* donné par l'opération : nombre de transferts réussis/nombre d'occasions de transférer, qui désigne la capacité du sujet à généraliser le principe de réussite appris au cours de l'aide.

➤ Utilisation

Cette épreuve, décrite ici comme illustration d'une démarche d'évaluation dynamique est principalement destinée à l'évaluation de l'intelligence de

sujets présentant des écarts et/ ou des carences socioculturelles par rapport aux populations habituellement évaluées. Elle permet en particulier, selon une approche principalement clinique, de tester l'hypothèse de déficit culturel.

Épreuve d'évaluation dynamique basée sur le SPM de Raven

L'épreuve, élaborée par Loarer et collaborateurs (*cf.* Loarer et Chartier, 1994 ; Loarer, 2001) est basée sur les progressives matrices de Raven (voir pour une présentation de cette épreuve le chapitre 4) et utilise la procédure T-A-R. Des aides portant sur la résolution de chaque type d'items du test ont été conçues en s'appuyant sur les travaux de Carpenter, Just, & Shell (1990), de Laroche (1956), de Raven (1981) et sur un travail complémentaire d'analyse des procédures de résolution de chaque item réalisé par les auteurs.

La passation est individuelle. Le sujet réalise une première fois le test puis se voit ensuite proposer un apprentissage sur les items auxquels il a échoué. Enfin, il passe le test une seconde fois. Les deux passations se font en temps limité. La figure 7.6 présente un exemple d'aide, correspondant à un principe de résolution utilisé dans l'épreuve de Raven. La passation dure donc le temps nécessaire pour les deux passations complètes du SPM et pour la phase d'apprentissage, soit environ 1 heure et 30 minutes.

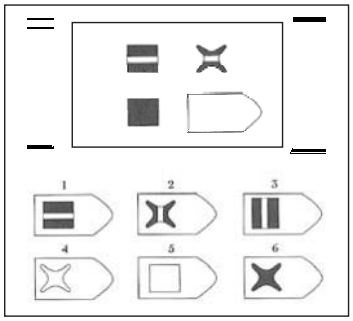

Figure 7.6
Exemple d'aide proposée correspondant à l'un des principes de résolution des items des Matrices de Raven SPM (Loarer et Chartier, 1994).

Une étude de validité de cette épreuve a été menée auprès d'élèves de 3e de collège (Loarer, 2001). La démarche a consisté à évaluer les élèves en début d'année scolaire à l'aide des matrices de Raven, selon la procédure T-A-R, et à comparer les résultats au pré-test (mesure classique) et les scores

au post-test (mesure incluant les effets de l'apprentissage) du point de vue de leur validité pronostique. Les notes scolaires aux 1r et 3e trimestres ont été prises comme critères (centrées et réduites par classe et par matière).

Les coefficients de validité ont été calculés initialement pour la totalité de l'effectif, puis une partition du groupe à la médiane a été opérée selon les résultats au pré-test. Les résultats montrent que pour l'ensemble des sujets, le post-test n'apporte pas, par rapport au pré-test, un surcroît significatif de validité. Par contre, après partition du groupe (*cf.* tableau 7.1) on constate que c'est le score au post-test qui est le plus valide pour les sujets ayant les scores les moins élevés, les coefficients de validité des scores au pré-test étant non-significatifs.

Tableau 7.1

Coefficients des corrélations (rbp) entre les scores aux SPM sans apprentissage (pré-test) et après apprentissage (post-test) et les notes scolaires aux 1er et 3e trimestres pour le groupe le plus faible

		Pré-test		Post-test
		r_{BP} sign.		r_{BP} sign.
Notes 1er trim.	09	ns.	.30	p<.05
Notes 3e trim.	.02	ns.	.22	p<.05

Nous retenons donc de cette étude que le score au retest après apprentissage reflète mieux le niveau réel des sujets les plus faibles.

En passation individuelle, l'interprétation du résultat au test à des fins de pronostic de la réussite ultérieure consiste alors à prendre le score au retest après apprentissage comme reflétant le niveau réel du sujet. La solution idéale serait de disposer d'un étalonnage des scores de retest pour différentes populations de référence. Il s'agit là d'une possibilité intéressante de développement de ce test.

Le Test d'Évaluation Dynamique de l'Éducabilité, 6e édition (T.E.D.E.6) de Pasquier

Le T.E.D.E. 6 a pour objectif principal, selon son auteur (*cf.* Manuel, p. 5), « la mesure de l'éducabilité de la personne par l'évaluation de son potentiel d'apprentissage ». II est destiné à des populations adultes, apprentis ou candidats apprentis sachant lire le Français.

Le test adopte la procédure apprentissage-test, variante de la procédure T-A-R sans effectuer le test initial. La mesure du potentiel d'apprentissage correspond à la performance obtenue par le test effectué après apprentissage.

➤ Tâche

Inspirée du test de calcul des longueurs de Faverge (1955), la tâche, de nature spatiale et logico-mathématique, est double : il s'agit d'une part de composer une égalité à partir de segments de droites et, d'autre part, de tracer les flèches figurant les superpositions de segments justifiant cette égalité (*cf.* figure 7.7)

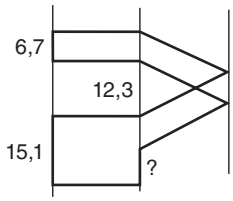

FIGURE C

2e EXEMPLE - Figure C.

On vous a donné 3 longueurs en trait fort : 6,7 ; 15,1 ; 12,3 et on vous demande de calculer une quatrième longueur en trait fort devant laquelle on a mis un point d'interrogation. Vous voyez quelle est la différence entre 15,1 - 6,7 - 8,4. La réponse est **8,4**.

Il était donc inutile de se servir de la longueur donnée 12,3.

Avez-vous bien compris ?

Figure 7.7
Exemple d'item du test de calcul des longueurs de Faverge (1955) et dont s'inspire le TEDE.

Le TEDE comprend trois niveaux de difficulté des items selon la complexité des opérations à effectuer.

Les 12 items de la phase d'apprentissage et les 18 items de la phase de test couvrent 3 niveaux de complexité des opérations à réaliser.

➤ Matériel et passation

Le matériel comprend deux livrets (le livret d'apprentissage et le livret de test) et un dossier d'instructions. La première phase de la passation est consacrée

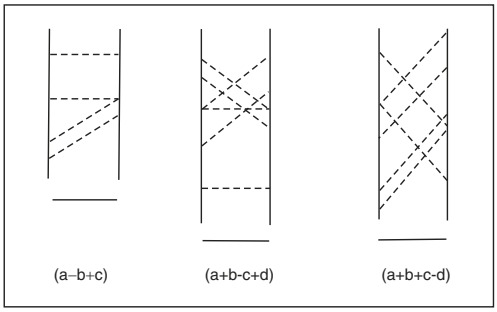

Figure 7.8

à l'apprentissage de la tâche à exécuter. Pour cette première phase, les sujets accompagnés de l'examinateur travaillent en situation d'auto-formation assistée à l'aide du dossier d'instructions et du livret pour l'apprentissage. Les personnes sont ensuite invitées à réaliser les exercices du livret de test (3 exemples + 3 séries de 6 items) sans aucune aide.

La passation peut être individuelle ou collective. Le temps de passation est de 2 heures (apprentissage : 1 heure + pause de 15 minutes + test : 45 minutes). Le temps de correction est d'environ 5 minutes (logiciel de correction).

➤ Éléments de validation

Le manuel présente de nombreuses données de validation, tant en ce qui concerne la validité interne du test (analyse des items, cohérence interne, analyse de biais différentiels, fidélité) qu'en ce qui concerne les validités externes (concourantes et prédictives). Ainsi, le manuel fait état d'une cohérence interne (coefficient alpha de Cronbach) de 0,92 et d'une stabilité temporelle (corrélation test-retest à une semaine d'intervalle de 0,94). Ces deux indices ont des niveaux très satisfaisants. On peut cependant regretter que la stabilité soit évaluée sur un échantillon très restreint (15 sujets). Les données relatives à la validité prédictive relativement à des indicateurs de réussite en formation s'appuient sur des échantillons plus larges (n=161 pour l'échantillon adulte et n=244 pour l'échantillon apprentis) et sont en moyenne élevées : les corrélations vont de 0,40 à 0,83 pour les différents groupes composant l'échantillon adultes et de 0,38 à 0,71 pour les différents groupes composant l'échantillon des apprentis.

Des éléments de validité concourante sont également présentés dans le manuel, relativement à différents tests de raisonnement et de connaissance, ce qui se comprend bien, mais aussi avec des inventaires d'intérêts (modèle de Holland) et de personnalité (épreuve de Gordon) ce qui se comprend moins aisément.

Nous retiendrons que le test présente globalement de bons indices de validité et semble bien adapté à la fois au type de populations visées mais aussi aux objectifs visés (prédiction de la réussite en formation).

Nous remarquerons cependant que le choix de la procédure d'« Aide-retest » ne permet pas de savoir quelle est la part du niveau initial et quelle est celle de l'apprentissage dans la réussite au test. La procédure permet probablement de limiter l'incidence de biais induits par le manque de familiarité avec la tâche, l'appartenance socioculturelle ou encore éventuellement l'émotivité. L'épreuve apparaît donc poursuivre principalement un objectif d'amélioration de la mesure de l'intelligence logique. Une analyse de la séquence d'apprentissage et de ces relations avec le test lui-même permet néanmoins d'aller bien au-delà de cet objectif et de fournir des indications relatives à l'apprentissage lui-même.

➤ Corrections et exploitation des résultats

Le test est fourni avec un programme permettant de saisir les résultats du sujet. Ce programme donne accès à différents traitements du protocole. Il permet de le positionner au regard de l'étalonnage adapté mais il permet aussi afin d'analyser les caractéristiques de la phase d'apprentissage et de dégager ce qui est appelé des « profils » du sujet. Il sont de trois types :

– profil fonctionnel,
– profil d'apprentissage,
– profil de transfert.

- *Le profil fonctionnel* correspond à la répartition des réussites selon les trois niveaux de complexité du test et selon les aspects de la tâche : égalités et déplacements ;
- *Le profil d'apprentissage* fournit une analyse des erreurs faites et de la façon dont elles ont été traitées ;
- *Le profil de transfert* reflète le lien entre la séquence d'apprentissage et le test proprement dit.

Cinq cas de figures ont été répertoriés : réussite réitérée (l'item est réussi à l'apprentissage et au test) ; gain (l'item non réussi à l'apprentissage l'est au test) ; perte (l'item réussi à l'apprentissage ne l'est plus au test) ; échec réitéré (l'item n'est réussi ni à l'apprentissage ni au test) ; omission réitérée (l'item n'est réalisé ni à l'apprentissage ni au test). Ces profils donnent accès à une information qui peut être utile dans une perspective psychopédagogique, ce que peu de tests permettent.

➤ **Étalonnages**

Les étalonnages disponibles portent sur une population d'adultes et sur une population d'apprentis.

Des étalonnages spécifiques sont en outre proposés :

– Pour la population adulte selon 4 niveaux de qualification :

groupe 1 : les hommes et femmes de niveau II et ID, les hommes de niveau IV, tous âges confondus,
groupe 2 : les hommes juniors (16- 29 ans) de niveau V et les femmes juniors (16-29 ans) de niveau IV,
groupe 3 : les femmes de niveau V, les hommes seniors (29-58 ans) de niveau V et les femmes seniors (29-58 ans) de niveau IV,
groupe 4 : les hommes et femmes de niveau VI, tous âges confondus ;

– Pour la population des apprentis et candidats apprentis, selon 4 niveaux de qualification des diplômes préparés : CAP, BEP, BP, Bac.

➤ **Restitution**

Le manuel fournit un modèle de fiche de synthèse et des indications concernant la restitution des résultats à la personne qui a passé l'épreuve. Cette restitution s'appuie en particulier sur l'analyse des profils qui a été faite.

9. Conclusions sur le potentiel d'apprentissage

La perspective de disposer d'instruments nouveaux permettant de dépasser les limites des instruments classiques est réjouissante. Pour cela, le concept d'évaluation dynamique constitue manifestement une « rupture épistémologique » (Paour *et al.*, 1995) par rapport aux approches traditionnelles de diagnostic

cognitif, et la démarche d'évaluation qui en découle semble prometteuse. L'évaluation dynamique apporte déjà une amélioration notable à la mesure de l'intelligence pour les sujets de faible niveau de performance. Neutralisant, au moins partiellement, les différences individuelles basées sur des inégalités socioculturelles, elle permet alors de réaliser un diagnostic plus valide (plus démocratique) de leurs capacités cognitives. Cependant, bien que très séduisante dans ses principes généraux, l'évaluation dynamique ne va pas sans poser un certain nombre de problèmes épineux. Comme le font remarquer très justement Büchel et Paour (1990), les tentatives pour dynamiser la psychométrie ont jusqu'ici soulevé autant de problèmes qu'elles en ont résolus.

D'un point de vue méthodologique, nous retenons les faibles qualités métrologiques des indices dynamiques et la bonne tenue des scores de post-test, tant du point de vue de la fidélité que de la validité. Nos observations confirment donc le choix fait par plusieurs auteurs de privilégier ce dernier (Guthke & Wigenfeld, 1992 ; Guthke *et al.*, 1997 ; Klauer, 1975, cité par Klauer, 1995 ; Speece, Cooper, & Kibler, 1990). Ce constat présente un intérêt évident pour le praticien. La prise en compte de la note au retest permet d'utiliser un indice dont la transparence facilite la restitution aux personnes concernées (sujets, enseignants,...) et dont la détermination évite la mise en œuvre d'un plan expérimental difficilement conciliable avec les contraintes des pratiques habituelles d'évaluation.

Pour ce qui est de ses finalités diagnostique et pronostique, trop de problèmes non résolus ou résolus seulement partiellement, notamment des problèmes relatifs à l'objectivité des observations, à la fidélité et à la validité des mesures, et à leur caractère plus ou moins analytique, subsistent. Le nombre de ces inconvénients, voire de ses limites, risque de la rendre difficile à mettre en œuvre dans la pratique courante des bilans psychologiques. Pour ce qui est de sa finalité éducative, il est probable que dès que l'on pourra proposer des méthodes d'évaluation fournissant de façon fiable des informations utiles aux formateurs et enseignants, bon nombre de praticiens se détourneront des méthodes statiques au profit de l'évaluation dynamique. Il s'agit donc d'une approche prometteuse et actuellement encore insuffisamment développée.

CHAPITRE
8

Utilisation des tests d'intelligence

Sommaire

QUI utilise des tests d'intelligence ? Comment les utilise-on ? Dans quels contextes ? C'est l'objet de ce chapitre.

Pour le « qui », les utilisateurs de tests sont, au moins en France, majoritairement des psychologues. Nous en exposerons les raisons.

Pour le « comment », nous détaillerons les grandes étapes de l'utilisation de tests psychologiques : de l'analyse de la demande à la restitution des résultats.

Enfin, concernant les contextes d'utilisation, ils sont nombreux : des bilans psychologiques effectués dans le système éducatif pour les scolaires (enfants et adolescents) aux pratiques de recrutement et de gestion des ressources humaines pour les adultes, sans oublier les pratiques plus contemporaines de conseil et d'accompagnement (bilan de compétences...), ou encore les bilans psychologiques effectués dans les hôpitaux (à la demande des psychiatres et les neurologues), sans oublier les demandes d'expertise des tribunaux... Nous présenterons les grandes lignes de quelques-uns de ces contextes d'utilisation.

1. Les conditions d'utilisation des tests

Qui peut utiliser des tests en France ?

Les utilisateurs de tests en France sont, comme nous venons de l'indiquer, le plus souvent des psychologues. En effet, un nombre important de tests n'est accessible qu'aux personnes pouvant justifier du titre de psychologue. Rappelons à ce propos que la profession de psychologue est réglementée depuis la loi de 1985. Pour pouvoir faire usage du titre de psychologue il est nécessaire : 1) d'avoir validé un cursus complet d'études supérieures en Psychologie (Licence et Master), 2) d'avoir réalisé (et validé) un stage dans un contexte professionnel d'une durée minimale de 500 heures.

La vente de tests psychologiques se trouve donc, du moins en France, et pour certains tests seulement, limitée aux personnes pouvant justifier du titre de psychologue. Pourtant, et la Société Française de Psychologie (SFP) le précise dans un document relatif à la « problématique de l'utilisation des tests » (SFP, non daté, disponible sur son site internet), du point de vue de la législation française actuelle, une personne non psychologue pourrait attaquer en justice un éditeur de test pour refus de vente. La restriction de

la vente relève donc plus d'un accord informel entre éditeurs et auteurs (et organisation professionnelle ?) que de l'existence de réels textes législatifs.

Certaines épreuves sont ainsi accessibles aux non psychologues, c'est par exemple le cas, pour rester dans le champ des tests de logique, des matrices de Raven, qui peut être considéré comme le, ou l'un des exemples prototypiques d'un test d'intelligence. L'accès libre à cette épreuve nous semble regrettable car le titre de psychologue constitue une garantie des capacités de l'utilisateur à utiliser de manière pertinente ce test.

Ce qui est en jeu n'est pas la défense d'un titre professionnel et de ses prérogatives, même si cet argument doit être pris en compte dans la réflexion sur le sujet, mais bien la protection des intérêts de la personne qui fait l'objet d'une évaluation. Le psychologue est généralement, du fait de sa formation et de son expérience, et de son code de déontologie, à même d'apprécier la pertinence d'utiliser ou non un test, de choisir le plus approprié à une situation donnée, d'estimer le niveau de fiabilité de l'épreuve sélectionnée compte tenu de la situation et du contexte, capable d'en interpréter correctement les résultats et de les restituer de façon adéquate à la personne... Bref, il possède les connaissances et compétences qui conditionnent un bon usage des tests.

Car il existe bien de mauvais usages des tests, et une personne non psychologue pourra être amenée, non pas nécessairement en raison de mauvaises intentions mais plus simplement par manque de connaissances et de vigilance sur certains aspects, à mettre en œuvre de telles pratiques néfastes, par exemple en utilisation mécaniquement le test, en l'interprétant sans nuance ou encore en l'appliquant hors de son champ de validité.

On peut argumenter qu'une grande diversité existe dans les formations de psychologie et que toutes ne fournissent pas de formation poussée en psychométrie. Cela est vrai et plus encore aujourd'hui depuis l'organisation des formations universitaires selon le système européen de formation (LMD) qui a abouti à augmenter la diversité des parcours de formation universitaires. Néanmoins, l'évaluation psychologique et la pratique des tests font partie des connaissances et compétences de base du psychologue et, même si tous les psychologues ne sont pas au sens strict du terme, spécialisés dans ce domaine, la formation qu'ils ont reçue et le code de déontologie qui encadre leur pratique constitue à ce jour la meilleure garantie en la matière.

Un article de Castro *et al.* (2001) est justement consacré à ce problème de l'utilisation des tests psychologiques par des psychologues et des non psychologues. À partir d'une enquête auprès de psychologues il ressort que ces derniers regrettent une absence totale de réglementation à ce

niveau et ne souhaitent pas que des non psychologues puissent utiliser des tests psychologiques : « l'ensemble des répondants s'oppose formellement à l'utilisation des tests psychologiques par des non-psychologues et ce pour deux raisons distinctes liées à la formation et à la notion de responsabilité » (p. 105). L'activité d'évaluation par des tests psychologiques est perçue comme un « acte psychologique » nécessitant un haut niveau de formation en psychologie. Les auteurs de l'article, dans leurs commentaires sur les résultats de l'enquête, avancent les arguments suivants :

> « L'utilisation déontologique des tests (dans l'intérêt des personnes évaluées) suppose donc un niveau de formation approprié, qui ne peut être atteint qu'à travers un cursus complet de psychologie » (p. 107).

Sept ans après ce constat, la réglementation n'a pas changé et les pratiques d'évaluation psychométriques par des non-psychologues a plutôt tendance à se développer. Plusieurs raisons peuvent être invoquées. Une raison évidente est de nature commerciale : alors que le « marché » de l'évaluation psychologique est actuellement en plein essor, en particulier sous l'effet du développement des tests informatisés et des tests en ligne, il peut apparaître souhaitable à certains (notamment à certains éditeurs) de laisser les choses en l'état. Une autre raison est peut-être à rechercher au sein même de la profession qui ne présente pas de réel consensus sur cette problématique.

Une pratique de diffusion des tests conditionnée au suivi d'une formation spécifique se développe actuellement. Elle concerne essentiellement mais non exclusivement les épreuves de type questionnaires (intérêts professionnels ou dimensions de la personnalité). Cette pratique consiste à conditionner la vente de l'épreuve, et donc son utilisation, au suivi obligatoire d'une formation courte relative à l'épreuve vendue. Même si cette pratique présente des limites (et constitue un vrai marché en soi car ces formations sont onéreuses) elle constitue à l'évidence un progrès par rapport à une situation de vente libre des tests. Cependant, de telles formations courtes ne peuvent être profitables qu'à des personnes ayant déjà des prérequis dans le domaine de l'évaluation psychologique. En outre, une application stricte de la règle aboutit parfois à obliger des psychologues parfaitement à même d'utiliser les épreuves à suivre également cette formation...

Une réflexion est donc à mener sur les conditions d'une ouverture de l'utilisation des tests à des non-psychologues. Quels aspects de la pratique et sous quelles conditions (d'expérience, de formation à la psychométrie, de formation spécifique à l'épreuve...) la pratique des tests pourrait-elle être élargie à des non-psychologues ? Cela peut probablement dépendre

du type de test et de l'expérience professionnelle[1] du « non-psychologue » qui souhaite utiliser une épreuve. Une contribution à cette réflexion est proposée par la SFP (SFP, non daté).

Si la passation d'une épreuve, ou du moins de certaines épreuves, voire même leur cotation, peut éventuellement faire l'objet d'une formation relativement limitée, il n'en est pas de même pour l'interprétation des résultats, la réflexion sur l'usage de tests dans le cas d'une pratique professionnelle, sur le respect de la personne etc. Tous ces éléments justifient bien le haut niveau de formation nécessaire pour pouvoir exercer des activités de psychologue, en lien direct avec les aspects éthiques et déontologiques de cette profession.

Afin de garantir la qualité de cette activité, d'indiquer quelles devraient être les éléments d'une bonne pratique professionnelle, mais également de garantir les droits des usagers, les organisations professionnelles de psychologues (dont la SFP) ont élaboré un code de déontologie afin de cerner les droits et les devoirs du psychologue, de définir un cadre de référence : « le présent code de déontologie est destiné à servir de règle professionnelle aux hommes et aux femmes qui possèdent le titre de psychologue, quels que soient leur mode d'exercice et leur cadre professionnel, y compris leurs activités d'enseignement et de recherche ».

Le code de déontologie des psychologues

La dernière version du code de déontologie date de 1996. Elle figure en annexe de cet ouvrage. Nous en reprendrons ici quelques articles, en lien direct avec l'utilisation des tests.

Au tout début de ce document, dans les principes généraux, il est indiqué que le psychologue décide du choix de ses méthodes :

> « Dans le cadre de ses compétences professionnelles, le psychologue décide du choix et de l'application des méthodes et techniques psychologiques qu'il conçoit et met en œuvre. Il répond donc personnellement de ses choix et des conséquences directes de ses actions et avis professionnels. »

Ces méthodes doivent reposer sur des fondements théoriques solides et explicites :

1. Que pourrait être une Validation des Acquis de l'Expérience sur ce point ?

> « Les modes d'intervention choisis par le psychologue doivent pouvoir faire l'objet d'une explicitation raisonnée de leurs fondements théoriques et de leur construction. Toute évaluation ou tout résultat doit pouvoir faire l'objet d'un débat contradictoire des professionnels entre eux. »

Ces éléments sont repris ensuite dans l'article 18 du code :

> « Les techniques utilisées par les psychologues pour l'évaluation, à des fins directes de diagnostic, d'orientation et de sélection, doivent avoir été scientifiquement validées. »

Le psychologue doit être capable d'estimer la fiabilité des mesures qu'il réalise mais aussi, plus globalement, des épreuves qu'il utilise, comme nous l'indiquent les articles suivants :

> « Le psychologue est averti du caractère relatif de ses évaluations et interprétations. Il ne tire pas de conclusions réductrices ou définitives sur les aptitudes ou la personnalité des individus, notamment lorsque ces conclusions peuvent avoir une influence directe sur leur existence » (article 19) ; « La pratique du psychologue ne se réduit pas aux méthodes et techniques qu'il met en œuvre. Elle est indissociable d'une appréciation critique et d'une mise en perspective théorique de ces techniques. » (article 17.)

Cette capacité de maîtrise des tests, outils et méthodes doit faire l'objet d'une formation spécifique dans laquelle :

> « Il est enseigné aux étudiants que les procédures psychologiques concernant l'évaluation des individus et des groupes requièrent la plus grande rigueur scientifique et éthique dans leur maniement (prudence, vérification) et leur utilisation (secret professionnel et devoir de réserve), et que les présentations de cas se font dans le respect de la liberté de consentir ou de refuser, de la dignité et du bien-être des personnes présentées. » (article 32.)

Mais certaines de ces règles de « bonne conduite » proposées dans le code de déontologie ont parfois quelques difficultés à être respectées dans les situations concrètes. Par exemple : qu'en est-il du choix de ses épreuves lorsque le psychologue ne peut disposer, au sein de sa structure, que d'un nombre parfois très limité de tests ? Qu'en est-il de la restitution des résultats dans le cadre d'utilisation de tests dans une procédure de sélection ?

L'élaboration d'un code est nécessaire mais faut-il encore qu'il soit facilement applicable et adapté aux différentes situations professionnelles. Pour Huteau et Lautrey « les indications fournies par les codes de déontologie

et les textes législatifs demeurent souvent assez vagues et leur application est parfois problématique » (1997, p. 110). Ils en donnent alors quelques exemples :

> « Qu'est-ce qu'une technique scientifiquement validée si l'on ne se réfère pas à des normes ? [...] La confidentialité est forcément mise à mal dans les procédures de recrutement : le psychologue est bien obligé de fournir des informations concernant les candidats puisqu'il est payé pour cela ! » (p. 110).

Le code de déontologie est un outil nécessaire et indispensable, c'est une référence pour une profession mais c'est au psychologue, en dernier recours, d'estimer, en fonction du contexte, quelle application de ce code est possible.

Claude Lévy-Leboyer, dans un article ancien, mais toujours pertinent, consacré aux problèmes éthiques posés par l'usage des tests (Lévy-Leboyer, 1987) distingue trois questions principales : « Le problème déontologique renvoie donc à trois questions qui sont, en fait, de nature méthodologique :

– quelles règles d'application faut-il respecter ?
– quelle est la valeur de l'outil que constituent les tests eux-mêmes ?
– comment la prouver de manière objective et réaliste à la fois ? » (Lévy-Leboyer, 1987, p. 473).

Ces trois questions, qui rejoignent certains points du code de déontologie relatifs à l'évaluation des personnes, peuvent guider le psychologue dans ses activités d'évaluation. Pour Lévy-Leboyer il existe aussi un lien étroit entre formation et compétences requises pour utiliser des tests dans de bonnes conditions : « seuls ceux qui ont reçu une formation théorique et pratique adéquate sont capables de choisir des tests adaptés à chaque situation, de les faire passer dans des conditions satisfaisantes, de les interpréter et de les utiliser dans le cadre plus large de décisions concernant la carrière des individus, et leur orientation, ou encore d'activités de conseil psychologique » (p. 474).

Un autre aspect intéressant de l'article concerne les décisions importantes qui peuvent être prises à partir des résultats de tests. Pour Levy-Leboyer il faut multiplier les sources d'information sur le sujet, et également, si possible, multiplier les personnes en charge de la décision : « d'une part, aucune décision ne peut être prise sur la base d'un seul test, ni même en fonction des seules informations que les tests apportent ; d'autre part, aucune décision importante ne devrait être prise par une seule personne » (Lévy-Leboyer, 1987, p. 484).

Le lecteur souhaitant approfondir la réflexion sur les aspects déontologiques de l'activité de psychologue pourra consulter les publications de Bourguignon (2000 et 2003) ainsi que le numéro spécial de janvier 2000 de la revue *Bulletin de psychologie* consacré à « Éthique en psychologie et déontologie des psychologues ».

Qui diffuse les tests en France ?

Les tests sont diffusés en France par des entreprises d'éditions. Historiquement en France, les plus anciennes, et sans doute les plus connues, sont les EAP[1] et les ECPA[2], regroupées depuis quelques années au sein des ECPA. On peut également signaler la présence, plus récente, d'autres éditeurs de tests tels qu'Eurotests, Hogrefe, OPP... Nous avons recensé en fin de chapitre les coordonnées des principaux éditeurs de tests en France. Le psychologue pourra consulter sur internet le catalogue de ces différents éditeurs et s'apercevra rapidement que certaines maisons d'édition, plus anciennes, possèdent un nombre important d'épreuves tandis que d'autres, plus petites et/ou plus récentes, ont un catalogue plus réduit et/ou en cours de développement. Signalons enfin que certaines de ces entreprises organisent des présentations de tests, et de nouveautés, sous forme de « petits-déjeuners ». C'est l'occasion, pour le psychologue, de se tenir informé de l'actualité des tests.

Les politiques de ces éditeurs peuvent être sensiblement différentes : certains sont plus spécialisés dans les outils à destinations des adultes, d'autres proposent également des tests à destination d'enfants ; certains tentent de diffuser des épreuves européennes et/ou francophones alors que d'autres adaptent surtout des tests d'origine anglo-saxonne.

Avant d'acheter un test il est fortement conseillé, si l'on ne connaît pas l'épreuve, de se rendre chez l'éditeur afin de pouvoir consulter l'épreuve dans son ensemble, et tout particulièrement les informations contenues dans le manuel qui accompagne le test.

1. Éditions et Applications Psychologiques.
2. Éditions du Centre de Psychologie Appliquée.

La formation à l'utilisation des tests

Nous pouvons distinguer ici trois types de formation offrant des enseignements sur la pratique des tests : les formations universitaires en Psychologie, les formations de psychologues à statut fonctionnaires et les organismes de type formation continue.

➤ Les formations universitaires en psychologie[1]

Comme nous l'avons indiqué auparavant, toute formation de psychologue doit comporter un enseignement significatif dans le domaine des tests. Généralement, une première approche de la mesure en psychologie, et des tests, est proposée aux étudiants pendant la Licence de Psychologie. Cette formation est ensuite développée en Master, mais restreinte au domaine spécifique de la spécialité du Master : par exemple, on ne présentera pas aux étudiants d'un Master de psychopathologie, les mêmes épreuves qu'aux étudiants suivant un Master en Psychologie du travail. De plus, le nombre limité d'heures de cours ne permet généralement pas d'aborder un ensemble vaste d'épreuves. Enfin, la place accordée à l'évaluation dans les programmes de formation peut dépendre du contexte historique et institutionnel de chaque Université.

Toutes ces sources possibles de variations expliquent que, même si tout étudiant diplômé en psychologie, de niveau Master, possède théoriquement les bases théoriques, méthodologiques et les compétences pratiques, nécessaires à la bonne utilisation de tests, il est possible que, comme dans la ferme des animaux (Orwell), certains soient plus « égaux » que d'autres à ce niveau. Et ces différences de formation auront des conséquences dans la pratique ultérieure, comme nous l'indique Dana Castro : « *deux choses sont certaines : toute la multitude de tests actuellement disponible n'est pas « enseignable » et les enseignements dispensés aux futurs psychologues au cours de leur formation initiale vont marquer de leur empreinte, et pour longtemps, leur pratique ultérieure* » (Castro, 2001, p. 52).

De plus, au-delà de la formation, une réelle expérience est indispensable :

> « Le fait de posséder des diplômes universitaires en psychologie ne remplace pas l'expérience acquise et tous les psychologues diplômés ne sont pas forcément compétents pour tous les tests existants. » (Levy-Leboyer, 1987 p. 474).

1. Nous pouvons intégrer également ici le cursus de psychologie du travail proposé par le CNAM.

➤ **Les formations de psychologues à statut de fonctionnaires**

Deux formations de psychologues amenés à exercer avec un statut de fonctionnaire dans l'éducation nationale sont évoquées ici : il s'agit de formations de type universitaire mais à recrutement particulier : les psychologues scolaires et les Conseillers d'orientation-psychologues.

Les psychologues scolaires interviennent dans l'enseignement primaire. Leur formation (d'une durée d'un an) est réservée aux instituteurs ou professeurs des écoles, titulaires d'une licence de psychologie. Les Conseillers d'orientation-psychologues (COP) interviennent dans les établissements d'enseignement secondaire et dans le supérieur, ainsi que dans les Centres d'Informations et d'Orientation (CIO). Le recrutement, sur concours, est ouvert aux titulaires d'une licence de psychologie et la formation dure ensuite deux ans. Dans ces deux formations des enseignements significatifs portent sur l'évaluation psychologique, les tests et l'examen individuel. Nous présenterons plus loin quelques éléments descriptifs des activités professionnelles des COP dans le domaine de l'évaluation des personnes.

➤ **La formation continue**

Le psychologue est tenu de maintenir ses connaissances à jour et doit pouvoir bénéficier de stages de formations. Les universités et des instituts spécialisés (comme par exemple l'INETOP), mais aussi des cabinets privés ou encore les éditeurs de tests, proposent des formations continues dans le domaine de l'évaluation psychologique. Il peut s'agir de formations portant sur des modèles théoriques, sur la pratique d'une épreuve ou d'un groupe d'épreuves (analyse de protocoles, études de cas...), ou encore de formations spécifiques accompagnant la sortie d'une nouvelle épreuve, ou d'une version rénovée (comme par exemple les formations sur le WISC-IV proposées par les ECPA et l'INETOP).

L'approche par la définition de normes et par l'analyse des compétences des utilisateurs de tests

Nous venons d'aborder les conditions d'utilisation des tests en France et avons signalé que cette possibilité d'utilisation reste marquée, dans ce pays, et pour différentes raisons, par la distinction entre psychologue et non psychologue. D'autres pays ont suivi une approche différente et se sont questionnés sur les compétences que devrait posséder tout utilisateur de

tests. Il faut signaler ici le travail important réalisé il y a quelques années par plusieurs organisations américaines de psychologues et professionnels de l'évaluation[1], repris et traduit en 2003 par Georges Sarrazin et collaborateurs (Sarrazin (Ed.), 2003). On peut noter que ce travail de traduction a été réalisé à l'initiative de l'Ordre de conseillers et conseillères d'orientation du Québec.

L'objectif de cet ouvrage est de proposer des normes de référence pour toute utilisation des outils d'évaluation :

> « L'objectif visé par les Normes de Pratiques est de promouvoir une utilisation valide et éthique des tests et de fournir une base à l'évaluation de la qualité des pratiques de testing. » (p. 1)

Il s'agit de proposer à la fois des critères d'évaluation pour les tests mais également des normes dans la pratique de ces instruments de mesure afin d'en garantir une bonne utilisation :

> « Pour être efficace, le testing et l'évaluation requièrent de tous ceux qui participent au processus la possession de connaissances, d'habiletés et d'aptitudes » (p. 2).

Sont ainsi visés les utilisateurs de tests mais également les concepteurs et éditeurs.

L'ouvrage est structuré en trois parties. Dans la première, *Construction de tests, évaluation et documentation,* sont abordées les principales notions psychométriques (qui ont été présentées dans le chapitre 2 de notre livre). La deuxième, *Équité en évaluation,* est relative à l'analyse de biais potentiels dans les tests[2]. La troisième partie, *Application du testing,* est consacrée aux conditions d'une bonne utilisation des tests.

Chaque partie est composée de plusieurs chapitres et à la fin de chaque chapitre figure une liste de normes. Prenons quelques exemples afin d'illustrer la démarche des auteurs :

Norme 1.2, relative à la validité des tests (partie I du livre) :

> « Les concepteurs de tests devraient expliquer clairement la façon d'interpréter et d'utiliser les scores d'un test. La ou les populations pour lesquelles le test a été conçu devraient être clairement délimitées et la

1. American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

2. L'équité étant définie comme « *une absence de biais ou le fait que tous les candidats sont traités également dans le processus d'évaluation* » (p. 90).

construction mentale que le test est censé mesurer devrait être décrite avec précision » (p. 19).

Norme 10.1, relative à l'évaluation des personnes présentant un handicap (partie II du livre sur l'équité) :

> « Dans l'évaluation des personnes handicapées, ceux qui conçoivent, administrent et utilisent les tests devraient prendre toutes les mesures nécessaires pour garantir que les inférences faites à partir des scores reflètent avec exactitude la construction mentale en cause, plutôt qu'un handicap ou les attributs qui lui sont associés sans rapport avec l'objet de la mesure. » (p. 127).

Norme 11.1, relative à la responsabilité des utilisateurs de tests (partie III du livre) :

> « Avant d'adopter et d'utiliser un test publié, son utilisateur devrait analyser et évaluer les documents fournis par son concepteur, particulièrement le matériel qui résume les objectifs du test, spécifie ses modalités administratives, définit les populations ciblées et passe en revue les possibles interprétations de scores basés sur des données fiables et fidèles. » (p. 136).

Comme nous pouvons le constater, ces normes visent bien l'ensemble du processus d'évaluation, de la qualité de l'instrument de mesure jusqu'aux connaissances et compétences que l'utilisateur de test doit maîtriser.

Tout particulièrement, c'est dans cette dernière partie de l'ouvrage (partie III) que l'on trouve les recommandations (et normes) relatives aux compétences et qualifications que doit posséder tout utilisateur de test. Ainsi, comme nous l'avons déjà souligné dans notre ouvrage, l'utilisateur de test doit être capable d'exercer un regard critique, un regard d'expert, sur le test qu'il souhaite utiliser :

> « Quand il sélectionne un test, le professionnel fait davantage que revoir le nom du test ; il fonde sa décision sur les preuves de validité et de fidélité et sur l'applicabilité des données normatives qui sont disponibles pour ce test dans la recension de la documentation de recherche. En plus tout à fait versé dans les procédures administratives appropriées, le professionnel doit aussi être familier avec les preuves de validité et de fidélité spécifiques à l'utilisation envisagée et avec les objectifs visés par les tests et les inventaires choisis et doit être prêt à développer une analyse logique soutenant les différentes facettes de l'évaluation et les inférences qui en découlent. » (p. 144).

Et c'est tout naturellement que ces aspects de l'activité professionnelle sont mis en relation avec les qualifications de l'utilisateur de tests, comme, par exemple, dans la norme 11.3 (p. 136) :

> « La responsabilité de l'utilisation d'un test devrait être uniquement assumée (ou déléguée) par des personnes formées à cette fin, possédant toutes les compétences professionnelles et l'expérience requise pour en prendre charge. Toute qualification particulière pour administrer ou interpréter le test et spécifiée dans le manuel devrait être respectée. »

Ce travail de réflexion, et de propositions de normes, très intéressant, ne semble pas être (très) connu en France. Un autre document international, et c'est l'objet de la partie suivante, a fait l'objet d'une plus large diffusion.

Les recommandations internationales sur l'utilisation des tests

La Commission Internationale des Tests (*International Test Commission*) a publié en 2000 des « Recommandations internationales sur l'utilisation des tests ». Ce document (31 pages) a fait l'objet d'une adaptation en langue Française, diffusée en juin 2003 par la SFP[1] (dans le cadre d'un numéro spécial hors série de la Revue Pratiques Psychologiques) et disponible sur son site internet (www.sfpsy.org). Nous ne présenterons ici que quelques extraits de ce document que le lecteur intéressé pourra trouver en annexe de cet ouvrage.

Ces recommandations ont été élaborées à partir de l'analyse de différents documents relatifs aux tests : « le travail sur les recommandations a débuté en rassemblant les documents se rapportant aux normes sur les tests, codes de déontologie, d'utilisation de tests, etc., dans un grand nombre de pays » (p. 10).

Les Recommandations comportent de 3 parties :

1) Les recommandations générales (p. 13-16) ;

2) Les recommandations concernant un usage éthique des tests (p. 17-18), [*agir de façon professionnelle et éthique, s'assurer de ses compétences, prendre ses responsabilités dans l'utilisation des tests, sécurité du matériel, confidentialité des résultats*] ;

3) Les recommandations pour assurer une pratique correcte dans l'utilisation des tests (p. 19-24) [*estimer l'intérêt éventuel d'une utilisation des tests*

1. Société Française de Psychologie, qui est l'une des associations professionnelle française de psychologues.

dans une situation d'évaluation donnée, choisir des tests techniquement fiables et appropriés à la situation, s'assurer de l'absence de biais, faire les préparations requises pour la séance de tests, administrer les tests de manière appropriée, corriger et analyser les tests avec exactitude, interpréter les résultats de manière appropriée, communiquer les résultats clairement et précisément aux personnes concernées, contrôler l'adéquation du test et de son utilisation].

Les objectifs de ce texte rejoignent les objectifs du travail sur les normes que nous venons de présenter. En effet, il s'agit de « promouvoir une bonne utilisation des tests et d'encourager des pratiques exemplaires dans le domaine de l'évaluation » (p. 9). Mais ici le but ultime n'est pas de définir des normes mais plutôt de lister les compétences que devrait posséder tout utilisateur de test : « le but à long terme de ce projet comprend la production d'un ensemble de recommandations qui se rapportent aux compétences (connaissances, capacités, savoir-faire et autres caractéristiques personnelles) requises des utilisateurs de tests. Ces compétences sont définies en terme de critères de performances évaluables. » (p. 10).

On trouve ainsi une liste de connaissances et compétences que devrait posséder tout utilisateur de test :

Connaissances déclaratives pertinentes

- Connaissances des principes et procédures de base de la psychométrie, et des exigences techniques des tests (par exemple, fidélité, validité, standardisation) ;
- Connaissance suffisante des tests et de la mesure, pour permettre une compréhension appropriée des résultats des tests ;
- Connaissance et compréhension des théories pertinentes et des modèles des aptitudes, de la personnalité et d'autres construits psychologiques ou de la psychopathologie, autant que nécessaire pour s'informer sur le choix des tests et l'interprétation des résultats ;
- Connaissance des tests et des fournisseurs de tests dans le secteur d'application où l'on intervient.

Connaissances pratiques et compétences

- Connaissances et compétences relatives aux procédures spécifiques d'évaluation ou aux instruments, y compris l'utilisation des procédures d'évaluation assistée par ordinateur ;
- Connaissances spécialisées et compétences pratiques nécessaires pour une bonne utilisation des tests situés à l'intérieur du répertoire d'outils d'évaluation de chacun ;

– Connaissances et compréhension de la – ou des – théorie(s) sous-jacente(s) aux scores au test, lorsque c'est important si l'on veut être en mesure de tirer des inférences valides à partir des résultats au test.

Ces Recommandations, fortes intéressantes aussi bien pour la pratique des tests que pour l'enseignement, et malgré leur diffusion par la SFP, semblent pourtant assez méconnues des praticiens...

Enfin, nous pouvons également citer une version de ces Recommandations concernant les tests informatisés et les tests sur internet disponible également sur le site de la SFP.

2. La pratique des tests

Quelques rappels

Avant d'aborder plus concrètement les principales étapes de la pratique de tests il nous semble nécessaire de rappeler un certain nombre de points importants, que nous avons déjà abordés dans les chapitres précédents mais que nous avons souhaité regrouper ici car ils conditionnent la fiabilité d'une pratique évaluative. Ils concernent : le manuel du test, l'erreur de mesure, la notion de biais, l'étalonnage, les tests informatisés et les tests en ligne.

➤ Le manuel du test

Comme nous l'avons déjà indiqué à plusieurs reprises, tout test doit être accompagné d'un, ou de plusieurs, manuel(s). La consultation du manuel est très importante et une première information sur la qualité probable du test pourra être inférée à partir de l'épaisseur de celui-ci : en effet certains manuels sont très minces alors que d'autres sont plus conséquents, avec parfois séparation en plusieurs volumes. C'est le cas par exemple du WISC-IV qui propose deux manuels : un manuel pour la passation et la cotation et un manuel pour les qualités psychométriques et l'interprétation des résultats (voir présentation de ce test dans le chapitre 3 de ce livre).

Que doit comporter un manuel ? Bien entendu le psychologue va y trouver toutes les indications utiles pour la passation et la cotation de l'épreuve (consignes, temps, matériel, étalonnages...). Il peut y trouver également des aides pour l'interprétation des résultats (comme par exemple des études de

cas). Enfin il doit y trouver toutes les études relatives à l'expérimentation de l'épreuve et à l'analyse de ses qualités psychométriques.

Le psychologue doit conserver un esprit critique sur les informations contenues dans les manuels et, par exemple, être capable de cerner les intérêts mais aussi les limites du test qu'il compte utiliser à partir de l'analyse de ces informations.

Il pourra éventuellement compléter les données du manuel par d'autres sources d'informations comme, par exemple, les ouvrages et publications spécialisés. On peut indiquer à ce propos la diffusion régulière d'un cahier « outils, méthodes et pratiques professionnelles en orientation » dans la revue *L'Orientation Scolaire et Professionnelle* destiné à présenter « un outil, une méthode ou une pratique d'aide à l'orientation » et dans lequel figure régulièrement la présentation de tests et/ou de pratiques d'évaluation.

➤ L'erreur de mesure

Il convient toujours de se rappeler que le score observé (la mesure) n'est qu'une estimation du score « vrai » du sujet. Comme nous l'avons indiqué, il est possible d'estimer cette erreur de mesure (le manuel comporte souvent une rubrique à ce propos), certains tests incitant même fortement le psychologue à encadrer chaque score obtenu d'un intervalle de confiance (voir par exemple les échelles de Wechsler).

Si l'on ne souhaite pas, pour différentes raisons, procéder à ce calcul, il faut, au minimum, prendre en compte l'erreur de mesure de manière plus qualitative dans l'analyse des résultats, en relativisant par exemple la caractérisation des performances du sujet par un seul score étalonné (et tout particulièrement lorsque le score brut du sujet est proche du seuil qui sépare deux scores étalonnés).

➤ Les biais

Bien que l'analyse des biais dans les tests soit de plus en plus fréquente, elle reste le plus souvent assez superficielle (Vrignaud, 2002a). Le psychologue devra toujours s'interroger sur les biais potentiels d'utilisation d'une épreuve sur un sujet, ou un groupe de sujets, particulier. Il sera, par exemple, attentif au vocabulaire contenu dans l'épreuve (est-il connu de tous les sujets ?), aux aspects culturels, et sociaux, qui pourraient avoir une influence, dans un sens comme dans l'autre, sur les résultats des sujets (connaissance a priori

de certains aspects du test ? familiarité avec la situation d'évaluation ?), aux modalités de présentation des items, et aux modalités de réponse...

➤ Les étalonnages

Rappelons ici qu'il est indispensable de s'interroger sur l'étalonnage, en particulier sur la date de recueil des données (effet Flynn), mais également sur la composition de l'échantillon des sujets de l'étalonnage (C.S.P, sexe...). Un examen minutieux de ces éléments permettra d'estimer dans quelles limites la comparaison des résultats d'un sujet avec la population de référence de l'étalonnage est adaptée.

Rappelons également qu'il est parfois possible d'obtenir des étalonnages supplémentaires (postérieurs à la publication du manuel par exemple) auprès de l'éditeur du test.

➤ Les tests informatisés et les tests en ligne sur internet

Depuis quelques années un véritable marché s'est ouvert dans le domaine des tests en ligne. Par exemple en entrant « test d'intelligence » sur un moteur de recherche, on obtient 383 000 réponses !!! Il ne s'agit pas bien entendu de 383 000 tests d'intelligence : dans un certain nombre de cas il s'agit de sites [1] qui proposent des passations gratuites de tests, ou d'un prix modique (quelques euros...), mais avec, le plus souvent, un « supplément » si l'on souhaite recevoir un compte rendu de la passation.

La qualité scientifique de ces tests est très variable. Il peut s'agir de tests semblables à ceux que l'on trouve dans des magazines, plus conçus pour distraire le lecteur que pour évaluer réellement ses carctéristiques. Ce sont alors des tests souvent très courts (peu d'items) et présentés de façon très attrayante. Ils témoignent souvent, comme l'observe Gaudron, d'une « imagination inversement proportionnelle à la validation scientifique » (Gaudron, 2008).

L'évaluation par les tests en ligne peut présenter une réelle solution pour les entreprises et les particuliers en raison de la souplesse et de l'économie de temps que représente ce mode de passation. Néanmoins, force est de constater qu'à ce jour, dans la plupart des cas, on ne dispose d'aucune information fiable sur les épreuves proposées, sur leur validité, sur les comparaisons

1. Il peut s'agir également de blogs personnels, et enfin, plus rarement, de liens avec de réels éditeurs de tests.

éventuellement effectuées avec un groupe de référence (étalonnages...). C'est l'un des points soulevés par Gaudron (1999 et 2008) qui liste un certain nombre de problèmes spécifiques posés par les procédures automatisées d'évaluation, dont les tests en ligne.

Pour les tests informatisés, en ligne ou non, les mêmes problèmes peuvent être soulevés. Il faut ici distinguer les versions informatisées de tests connus, déjà diffusés par des éditeurs de tests, des tests proposés par des entreprises et/ou des cabinets et/ou sur internet dont on ignore, bien souvent, l'origine.

Il faut rappeler par exemple, qu'un test informatisé doit présenter les mêmes qualités qu'un test « papier-crayon » et donc être accompagné d'un manuel... ce qui n'est pas toujours le cas dans certaines épreuves informatisées utilisées ou utilisables, par exemple, dans le domaine de la gestion des ressources humaines (qui reste l'un des grands « marchés » actuels de l'évaluation...). On ne peut qu'inciter le psychologue à une grande prudence dans ce domaine.

Il doit exercer son regard critique sur les informations communiquées relatives aux qualités psychométriques du test (et a fortiori sur leur absence !) afin d'estimer la fiabilité de l'épreuve qu'on lui propose.

Un document diffusé sur le site de la SFP et relatif aux « Recommandations Internationales sur les tests informatisés ou les tests distribués par internet » peut être d'une grande aide sur le sujet.

En cas d'interrogation sur la fiabilité d'un test informatisé, et/ou en ligne, le psychologue peut également consulter la commission des tests de la SFP.

La pratique des tests : de l'analyse de la demande à la restitution des résultats

Dans une pratique d'évaluation, nous pouvons distinguer plusieurs étapes entre la phase de réflexion sur la demande jusqu'à la phase de restitution des résultats :

1. Réflexion sur la demande et sur la pertinence d'utiliser des épreuves standardisées
2. Choix les épreuves adaptées,
3. Entretien préalable avec le sujet,
4. Passation des épreuves,
5. Cotation,
6. Interprétation des scores,
7. Préparation de la restitution,

8. La restitution orale,
9. La restitution écrite (le compte rendu).

Bien entendu cette décomposition en 9 étapes n'est qu'une des possibilités de rendre compte des différentes phases d'une pratique évaluative et doit être adaptée au contexte de l'évaluation (situation de sélection, de conseil, d'expertise...). Le plus important, quel que soit le nombre des étapes, est de prendre le temps d'aborder tous ces aspects.

➤ Réflexion sur la demande et sur la pertinence d'utiliser des épreuves standardisées

Rappelons que l'utilisation de tests se situe dans une pratique globale de psychologue et qu'il est nécessaire, avant toute intervention, d'analyser la situation. C'est l'une des distinctions probables entre une pratique (évaluative) d'un psychologue (pour qui l'analyse de la demande est essentielle) et une pratique évaluative d'un non psychologue (qui aura tendance à répondre directement à la demande exprimée). Pour tenter de comprendre tous les éléments de la demande le psychologue peut (doit) se poser les questions suivantes : Quelle demande est exprimée ? Exprimée par qui ? Dans quels termes ? Repérer la demande explicite et l'éventuelle demande implicite, la demande institutionnelle... Le plus souvent la simple question « À qui devrai-je présenter les résultats ? », permet de repérer le réel demandeur !

Le psychologue peut ainsi séparer la commande de la demande : « un premier niveau d'analyse doit s'appuyer sur la distinction classique des psychosociologues entre la *commande*, le besoin qui est explicitement exprimé, et la *demande*, qui pose le véritable problème et dont la personne a plus ou moins conscience » (Guillevic et Vautier, 1998, p. 19*).*

La (ou les) demande(s) étant précisée(s), et éclaircie(s), le psychologue doit alors se positionner, en tant que professionnel par rapport à cette demande : dans quelle mesure peut-il y répondre ? Comment ? Dans quelles limites ? Il peut se référer ici, si nécessaire, au code de déontologie.

Puis il doit cerner l'intérêt d'utiliser des épreuves standardisées pour répondre au problème posé : que vont apporter de plus, et/ou de spécifique, ce ou ces épreuves ? Et quels éléments du problème ne seront éventuellement pas pris en compte par ces épreuves ? Il s'agit bien ici de cerner les limites de l'utilisation de tests.

Cette phase d'analyse de la demande est essentielle car elle permet, bien souvent, de révéler la complexité d'un problème présenté comme une simple évaluation.

➤ Choisir les épreuves adaptées

Ce n'est qu'une fois que le problème sera correctement posé, et que l'une des solutions envisagées consistera à utiliser une ou plusieurs épreuves, qu'il faudra déterminer le(s) type(s) d'épreuve et leur nombre. La, ou les, épreuve(s) sera/seront adaptée(s) 1° au problème posé, 2° aux caractéristiques du sujet[1] (âge, sexe, niveau d'étude...).

Préalables : la connaissance des épreuves existantes et leur possibilité d'accès. Par exemple, si le psychologue ne dispose dans sa structure que d'un nombre limité d'épreuves, cet élément limite de fait le champ des possibles.

Le psychologue questionnera également son degré de connaissance, et de maîtrise, de ou des épreuves envisagées.

➤ Entretien préalable avec le sujet

Toute passation de tests doit être précédée d'un entretien avec le sujet. Le psychologue va ainsi s'assurer que la personne est bien consentante pour une passation de test (en particulier lors d'une procédure de recrutement...) et va recueillir des informations sur l'expérience éventuelle du sujet dans ce domaine (a-t-il déjà passé des tests ? dans quelles conditions ? ...).

Lors de cet entretien seront également évoqués les objectifs de la passation, les modalités de passation, le type de tâche à résoudre...

Le psychologue devra être conscient que la majorité des personnes a souvent beaucoup d'appréhension par rapport aux tests[2] et fera en sorte de préparer du mieux possible le sujet à la passation (dans les limites définies dans le manuel du test). Bernaud parle ainsi de pratiques de « testage ouvert » qui consiste à « *fournir aux participants, quelque temps avant la séance d'évaluation, des informations sur son contenu et les moyens de s'y préparer* » (2007, p. 87).

On peut rappeler l'existence d'un dispositif, assez rare, que l'on trouve dans la version APM des Matrices Raven : la possibilité de proposer au sujet

1. Il faut toujours s'interroger dès cette étape sur les étalonnages disponibles.
2. Et d'autant plus quand il y a des enjeux, comme par exemple, l'accès à un emploi ou à une formation...

une série d'items (un livret d'apprentissage), série non évaluée qui servira de préparation à la passation de l'épreuve proprement dite (voir dans le chapitre 4 de ce livre la présentation de la version APM des Matrices de Raven).

Si l'on pousse ce dispositif à l'extrême, nous nous retrouvons dans des situations proches de celles proposées dans le cadre d'une évaluation dynamique (voir chapitre 7 de ce livre).

➤ Passation des épreuves

Avant la passation, le psychologue prépare le matériel nécessaire, en quantité suffisante (en cas de passation collective...).

Lors de la passation des exemples, il est attentif aux éventuels problèmes de compréhension des consignes, de report des réponses du sujet... Il s'agit de créer ce que l'on désigne par une *relation positive* (Bernaud, 2007, p. 88), propice au bon déroulement du test et à l'implication du sujet : qualité de l'accueil du sujet, information claire et objective sur les procédures...

Pour la passation du test, le psychologue respecte scrupuleusement les consignes du manuel (le matériel, l'attitude du psychologue, les consignes, les exemples, le temps de passation...) de manière à garantir la standardisation. Tout en restant dans les limites de celle-ci, il doit s'assurer de la bonne compréhension, par le sujet, de la situation globale de la passation.

Il observera**,** si possible, la conduite du sujet pendant la passation : hésitations, lassitude, niveau de motivation, implication[1], découragement, comportement face à la difficulté (surtout dans les items difficiles), rapidité globale d'exécution, temps de passation... Ces observations seront plus faciles à réaliser dans le cadre d'une passation individuelle ou en petit groupe. Elles seront éventuellement à reprendre avec le sujet dans la phase de la restitution des résultats *(« il m'a semblé que vous avez hésité à tel moment... que vous vous êtes un peu découragé en fin d'épreuve... avez-vous manqué de temps ? »*) et pourront également être utiles dans la phase d'interprétation des résultats.

1. Essayer de repérer les réponses visiblement données au hasard (par exemple : cochage systématique des réponses en ligne ou en colonne...).

➤ Cotation des épreuves

Calcul des scores bruts

Pour la cotation du protocole des réponses du sujet, il faut suivre les indications du manuel. On accorde généralement 1 point pour chaque bonne réponse (à vérifier dans le manuel).

Au préalable le psychologue aura analysé la validité du protocole en vérifiant :

– l'absence de réponses systématiques au hasard (par exemple des réponses situées systématiquement en ligne ou en colonne...) ;
– la concordance entre réponses du sujet et items : vérifier que le sujet ne s'est pas trompé dans le report de ses réponses (attention aux décalages éventuels des réponses du sujet par rapport aux items...). En cas de doutes il est conseillé de reprendre quelques items à l'oral afin de vérifier l'absence de biais à ce niveau ;
– les éventuelles observations du sujet pendant la passation (implication...).

Les scores étalonnés

Le psychologue sélectionne le, ou les, étalonnages les plus adaptés et transforme les scores bruts en scores étalonnés.

Il doit être attentif à l'erreur de mesure, tout particulièrement quand le score brut du sujet se situe à proximité d'un seuil qui sépare deux catégories de notes étalonnées (dans ce cas il peut être préférable de situer le sujet sur ces 2 scores étalonnés).

➤ Interprétation des scores

Avant d'interpréter les scores étalonnés le psychologue doit analyser le ou les étalonnages disponibles (description de l'échantillon d'étalonnage, date de l'étalonnage, répartition selon le sexe, la CSP...) de manière à vérifier la fiabilité des comparaisons, et des éventuelles généralisations, qu'il va effectuer. Il convient toujours de limiter les conclusions sur le niveau de performance d'un sujet aux caractéristiques de la population d'étalonnage. Par exemple, si l'on observe de bons résultats à la batterie NV7, il faut considérer que ce ne sont pas des bons résultats dans l'*absolu* mais de bons résultats *relatifs* aux caractéristiques de la population d'étalonnage, qui est ici peu qualifiée...

Si l'on dispose de plusieurs scores, l'interprétation suit généralement le principe suivant : partir des scores les plus généraux (le score total) et aborder

ensuite les scores les plus spécifiques (notes aux différentes sous échelles, voire scores à certains subtests).

Qu'est-ce que chaque score représente ? Le manuel doit fournir des informations sur ce point. Il faut ici revenir vers le modèle théorique de référence (par exemple le facteur g s'il s'agit d'un test de ce type) afin de situer la performance observée dans un cadre théorique plus large. Il faut également prendre en compte les éventuelles spécificités de l'épreuve (type d'items, type de réponse, temps libre ou limité...) qui donnent une « coloration » de ce qui est plus précisément évalué par l'épreuve utilisée. Il peut être intéressant à ce propos de distinguer la *compétence* (ce que l'on cherche à mesurer) de la *performance* (mesure réalisée dans un contexte précis, avec un certain test...) pour aborder les limites de la généralisation de ce qui a été évalué.

Enfin, il faut mettre en relation les résultats observés et ce que l'on connaît, par ailleurs, du sujet (expérience, qualification, métiers exercés, projets...) : quels sont les résultats concordants ? Les résultats discordants ? Quels sont les éventuels décalages ?...

Analyse des erreurs ?

Il peut être parfois utile de procéder à une analyse des erreurs en repérant quels sont les items échoués et en essayant d'en comprendre la cause. On peut ainsi envisager de revenir sur ces échecs lors de l'entretien de restitution, afin de tenter de mieux comprendre le raisonnement du sujet.

On peut également distinguer l'absence de réponse d'une réponse fausse. Il peut s'agir également de repérer les absences de réponse en distinguant celles situées en cours d'épreuve (assimilables à un échec) de celles situées en fin d'épreuve (attribuables, au moins en partie, à un manque de temps dans le cas d'épreuve à temps limité). Il peut être intéressant par exemple de confronter le sujet aux items qu'il n'a pas eu le temps d'aborder lors de la passation de l'épreuve afin d'estimer sa capacité à résoudre les items situés en fin d'épreuve, qui sont souvent les plus difficiles. Bien entendu, on ne tiendra pas compte de ses éventuelles réussites supplémentaires dans son score, car elles se situent hors limites de temps, mais ces informations peuvent être utiles. Rappelons à ce propos que certaines épreuves proposent des étalonnages avec passation en temps libre (par exemple les Matrices de Raven).

➤ Préparation de la restitution

> « Avant de s'engager dans la phase de restitution proprement dite, un préalable indispensable consiste à :
> – s'assurer que le bénéficiaire a passé les épreuves dans de bonnes conditions ;
> – l'interroger sur l'intérêt induit par les épreuves ;
> – concevoir et proposer des méthodes susceptibles de l'aider à comprendre les résultats, à se les approprier, à y réagir de façon argumentée et, enfin, à en faire la synthèse. » (Blanchard *et al.,* 1999, p. 287.)

Il s'agit donc, avant de communiquer les résultats au sujet, de vérifier certaines conditions de validité (condition de passation, niveau d'implication...) mais aussi, et c'est l'objet de cette partie, de préparer la restitution en fonction des objectifs proposés par ces auteurs et en particulier de faire en sorte que le sujet 1) comprenne ses résultats, 2) se les approprie. Cette approche de la présentation des résultats est assez récente et concerne principalement l'utilisation de tests dans le cas de démarche de conseil et/ou de bilans. Dans ces contextes d'utilisation, l'importance de cette phase de restitution (on parle aussi de rétroaction) s'est considérablement développée ces dernières années. En effet, avec le développement des pratiques de conseil, l'objectif principal des évaluations s'est progressivement modifié et vise maintenant, en totalité ou en grande partie, à améliorer la connaissance de soi des individus. C'est le cas, par exemple, dans les pratiques de bilans de compétences où l'on cherche alors à développer la connaissance du sujet sur ses aptitudes, ses intérêts, ses traits de personnalité... L'objectif final étant de lui permettre de mieux faire ses choix d'orientation.

D'où un intérêt de plus en plus vif, dans les pratiques comme dans les recherches, pour cette phase de l'évaluation. Mais dans la majorité des cas cet intérêt se porte sur les questionnaires de personnalité ou de choix d'activités professionnelles (les questionnaires d'intérêts) et non sur les tests d'intelligence. On trouve, par exemple, dans l'ouvrage de Bernaud et Vrignaud de 2005, consacré à l'évaluation des intérêts professionnels, une présentation de plusieurs méthodes de restitution des résultats.

Par contre, peu d'études ou de recherches portent sur la restitution des résultats des tests d'intelligence logique. Pourquoi ? L'une des pistes explicatives tient peut-être au fait que les questionnaires d'intérêt sont plus fréquemment utilisés dans les pratiques de conseil que ne le sont les tests d'intelligence. Peut-être aussi qu'il y a plus d'éléments à restituer dans un tel questionnaire que dans une épreuve de performance.

Néanmoins, certains manuels donnent des indications sur cette phase de restitution. C'est par exemple le cas de la batterie NV7 (voir la présentation de ce test dans le chapitre 5 de ce livre). Le psychologue pourra donc s'inspirer des éventuelles informations des manuels pour préparer la restitution.

Nous allons présenter ici quelques éléments généraux concernant la restitution des résultats. Ces éléments sont particulièrement adaptés à une pratique d'évaluation destinée à accompagner le sujet dans une démarche de réflexion sur lui-même.

Quelle que soit l'épreuve, dans la phase de préparation de la restitution le psychologue doit se poser les questions suivantes :

Quoi dire ?

Cerner l'essentiel, et le superflu en fonction, d'une part, des capacités du sujet à prendre en compte ces informations, d'autre part, des objectifs de l'évaluation.

Comment le dire ?

Moduler les modalités de restitution, et le vocabulaire utilisé, en fonction des caractéristiques du sujet.

Comment faciliter la compréhension des résultats par le sujet ?

Il faudra éventuellement envisager de revenir sur l'épreuve (ce qu'elle mesure ? comment elle le mesure ?), en reprenant éventuellement des exemples d'items. Préciser également ce que le test ne prend pas en compte (ses limites).

Dans la mesure du possible il peut être pertinent de trouver (ou de faire trouver) des liens entre ce qui est évalué par le test et des situations de la vie quotidienne et/ou professionnelle du sujet.

Enfin, on peut inciter le sujet à reformuler avec ses propres mots les points les plus importants qui ont été abordés dans la restitution (l'inciter à prendre des notes par exemple...).

Comment aider le bénéficiaire de l'évaluation à gérer cette nouvelle information sur lui-même ?

Le résultat à un test d'intelligence logique, surtout lorsqu'il a donné lieu au calcul d'un QI, présente, qu'on le veuille où non, des enjeux de comparaison sociale. Il convient de prendre en compte cette dimension et d'aider le sujet à bien la gérer. Cela dépasse à proprement parler la phase d'évaluation et de compréhension de celle-ci par le sujet, et concerne les conséquences pour lui-même de cette évaluation. Un accompagnement de la personne sur ce point peut éventuellement nécessiter un rendez-vous ultérieur.

Le point essentiel ici est le suivant : il faut que la restitution soit adaptée au sujet, à ses caractéristiques personnelles (ses capacités de compréhension, son

niveau de langage...), à l'objectif de l'évaluation (pourquoi a-t-il demandé ou accepté de passer ces tests ? quelles sont ses attentes ?), ainsi qu'à sa situation personnelle.

Il faut également que la restitution soit adaptée au psychologue : chacun a son propre style, ou doit le trouver. Par exemple, certains psychologues peuvent utiliser des schémas, des illustrations à partir de courbes de Gauss (par exemple pour faire comprendre la notion d'étalonnage...) tandis que d'autres seront plus à l'aise dans des explications verbales...

On ne peut que conseiller au psychologue de rechercher (et de trouver) son style, de créer ses propres outils de restitution en élaborant, par exemple, quand il n'existe pas, un cahier (ou livret) de restitution pour le sujet qu'il complétera lui-même au fur et à mesure de la présentation de ses résultats.

Le psychologue doit laisser la place, dans cette phase de restitution, à l'expression par le sujet du « vécu » de sa passation : a-t-il été surpris par l'épreuve ? A-t-il rencontré des difficultés ? Que peut-il dire, après coup, de cette situation de passation ?

Dans la mesure du possible on envisagera une restitution « dynamique », pour le sujet, dans laquelle il pourra commenter les résultats présentés par le psychologue, les questionner voire les contester. En effet, « il est souvent plus éclairant et plus dynamisant d'inviter le bénéficiaire à commenter et à étayer ses résultats, voire à s'opposer à ceux-ci lorsqu'ils semblent contredire un aspect de sa personnalité ou de ses compétences » (Bernaud, 2000, p. 101).

La situation de restitution doit être conçue comme une situation de communication, d'échange (Guédon et Savard, 2000) et non pas comme (uniquement) un discours d'expert.

Enfin, il faut toujours prévoir à l'avance les éventuelles traces, les éventuels documents, que le sujet va (peut) conserver de sa passation et de ses résultats.

Une règle générale consiste à ne jamais laisser la feuille de passation au sujet. En effet, pour des motifs de respect des règles de copyright mais également pour des raisons déontologiques, il n'est pas possible que le sujet reparte avec un exemplaire du test. Par contre, plusieurs possibilités de conservation des traces de ses résultats sont possibles allant des simples notes prises par le sujet lui-même au cours de l'entretien au cahier de restitution, disponible avec certains tests (ou éventuellement élaboré par le psychologue).

➤ Le déroulement de la restitution orale

La restitution est le plus souvent individuelle mais on peut envisager, au moins à certains moments, une phase collective (en petit groupe).

La restitution doit se dérouler comme le psychologue l'a prévu (*cf.* la partie précédente) tout en s'adaptant aux réactions du sujet : une certaine souplesse est nécessaire !

Elle peut suivre le plan suivant :

1. Rappel des objectifs de l'évaluation,
2. Présentation des bases théoriques de l'épreuve utilisée : ce qu'elle mesure ; comment elle le mesure (rappels des caractéristiques des items...) ; les indicateurs qu'elle permet de calculer (le ou les scores)...
3. Présentation de la notion d'étalonnage et de comparaison à un groupe de référence,
4. Échanges sur les conditions de passation
5. Recueil éventuel des représentations a priori du sujet par rapport à ses résultats : comment pense-t-il qu'il va se positionner ?
6. Présentation des résultats
7. Confrontation entre les résultats du test et les représentations a priori du sujet
8. Bilan de l'évaluation par rapport aux objectifs de départ et analyse de ses conséquences.

On peut terminer l'entretien de restitution en demandant au sujet ce qu'il a retenu de l'entretien, ce qui lui semble le plus important, et/ou le plus surprenant... On pourra ainsi repérer quelles informations il a retenu (sélectionné) et quelles informations ne lui semblent pas essentielles... On peut aussi se rendre compte des éventuelles erreurs d'interprétation des résultats, des points à reprendre avec lui, de suite ou dans le cadre d'un autre rendez-vous.

Enfin, il convient de replacer l'évaluation dans la problématique générale du sujet : le test comme l'une des étapes d'un processus d'accompagnement.

➤ La restitution écrite (le compte rendu)

Dans la plupart des cas, le psychologue va rédiger un compte rendu écrit. La première question à se poser concerne le destinataire de ce document : est-ce un document de travail pour le psychologue qui ne sera pas communiqué à un autrui ? Est-ce un document pour le principal intéressé ? Pour une institution ? Pour un tiers à l'origine de la demande ? Pour un autre psychologue ?...Qui demande un compte rendu écrit, et pourquoi ?

Le psychologue peut ainsi être amené à rédiger plusieurs comptes rendus différents, plus ou moins étoffé, d'un même « cas » en fonction des destinataires, en adaptant à la fois la forme et le fond. Il doit également préciser, au début du document, quel est le destinataire et quelles sont les limites de diffusion de ce document.

Enfin, il doit dater et signer tout document qu'il est amené à rédiger.

L'un des règles du psychologue étant le respect de la confidentialité, le praticien devra faire preuve de prudence et de réflexion dans ses écrits. On peut retrouver cette règle de base dans le code de déontologie des psychologues : « les documents émanant d'un psychologue (attestation, bilan, certificat, courrier, rapport...) portent son nom, l'identification de sa fonction ainsi que ses coordonnées professionnelles, sa signature et la mention précise du destinataire. Le psychologue n'accepte pas que d'autres que lui-même modifient, signent ou annulent les documents relevant de son activité professionnelle. Il n'accepte pas que ses comptes rendus soient transmis sans son accord explicite, et il fait respecter la confidentialité de son courrier » (chapitre 2, article 14).

Dans le cas d'un document à usage interne, le psychologue réunit les différentes informations concernant le sujet pour élaborer un « dossier d'examen psychologique » de la personne. Dana Castro rappelle que ce dossier est la propriété du psychologue (Castro, 2006, p. 473).

Pour les autres types de comptes rendus, le contenu peut éventuellement être discuté avec le sujet : quelles informations nous a-t-il confiées (par exemple dans la phase d'analyse des résultats ou lors d'un entretien...) et qu'il ne souhaite communiquer à autrui ?

Quelles informations communiquer dans un compte rendu écrit ?

Comme nous venons de l'indiquer, ces informations vont dépendre du destinataire mais on peut préciser ici ce qui nous semble être la base de tout compte rendu écrit :

- Indiquer le ou les objectifs de l'évaluation (clarification de la demande) ;
- Préciser toujours les épreuves utilisées (et les situer rapidement dans un cadre théorique... avec les limites éventuelles de ce cadre...), les étalonnages consultés (date d'élaboration et caractéristiques des populations d'étalonnage...) et les éventuelles limites de comparaison compte tenu des caractéristiques de la personne évaluée ;
- Situer les résultats du sujet (sans oublier la notion d'erreur de mesure) ;
- Nuancer éventuellement ces résultats en fonction d'indications cliniques relevées durant la passation (motivation, arrêts, comportement face

à la difficulté...) et/ou d'informations sur le sujet (exemple : niveau d'expérience par rapport aux tests ...).

Terminer le compte rendu en articulant les résultats aux autres informations recueillies sur le sujet (résultats plus ou moins en accord avec ce qui était attendu en fonction, par exemple, du niveau d'étude du sujet et/ou de son expérience professionnelle...) et en analysant de quelle manière les résultats répondent à l'objectif de départ.

Enfin, il peut être pertinent de dire quelques mots sur la restitution orale : comment le sujet a-t-il réagi ? Ses remarques, son degré d'accord ou de désaccord avec les résultats, avec les interprétations proposées ? Les limites éventuelles de compréhension de ses résultats ?...

3. Exemples de contextes d'utilisation des tests d'intelligence logique

Comme nous l'avons indiqué dans l'introduction de ce chapitre, les tests d'intelligence sont utilisés dans plusieurs secteurs d'activité : l'école, l'entreprise, l'hôpital, le tribunal... Il est impossible de dresser un panorama exhaustif de ces contextes d'usages. Nous ne présenterons donc ici que certains d'entre eux qui nous semblent être les plus représentatifs : le système éducatif, le recrutement et les ressources humaines, les pratiques plus actuelles de conseil tout au long de la vie, et enfin, le secteur de la formation.

Dans le système éducatif

Dans l'enseignement public, il existe les psychologues scolaires, qui interviennent auprès d'élèves scolarisés dans l'enseignement élémentaire, et les Conseillers d'orientation-psychologue, qui interviennent principalement dans l'enseignement secondaire.

Les activités des psychologues scolaires[1] concernent un public d'enfants et ne correspondent donc pas à l'objet principal de cet ouvrage relatif aux tests utilisables auprès d'adolescents et d'adultes. Par contre, nous évoquerons

1. Pour une présentation des psychologues scolaires le lecteur peut consulter Cognet (2006).

les activités des conseillers d'orientation-psychologues[1] (C.O.P) qui sont amenés à utiliser des tests d'intelligence auprès d'adolescents et d'adultes.

Il s'agit principalement des situations suivantes :

1. Dans le cadre de leur activité d'aide à l'orientation, de conseil auprès de publics scolaires, ils peuvent utiliser des tests d'intelligence logique comme un **outil d'aide à la connaissance de soi** : identifier ses points forts, ses atouts... mais aussi ses faiblesses... Le COP peut utiliser ici des tests collectifs comme les batteries factorielles (rappelons que la DAT5 propose des étalonnages pour les collégiens et les lycéens) mais également des épreuves de type facteur g.
2. Alertés par les équipes pédagogiques des établissements scolaires au sujet d'élèves en difficulté, ils peuvent également procéder à une **évaluation des capacités cognitives, complémentaire aux évaluations scolaires**. C'est le cas, par exemple, d'élèves de 6e en grande difficulté scolaire et pour lesquels, après le plus souvent un redoublement, l'équipe pédagogique envisage une orientation vers un enseignement adapté. La loi de 2005 sur le handicap a modifié la procédure d'orientation vers ce type d'enseignement. Un dossier doit être transmis à une commission départementale d'orientation, dossier devant comporter un « bilan psychologique » établi par le C.O.P (circulaire n° 2006-139 relative aux enseignements adaptés dans le second degré). Ce bilan, précise la circulaire, doit être « étayé explicitement par des évaluations psychométriques ». Le plus souvent les COP utilisent alors une échelle d'intelligence, comme le WISC, qui permet l'estimation d'un Q.I. En deçà d'une certaine valeur de Q.I (généralement autour de 70/75), et en fonction d'autres informations complémentaires, le COP peut proposer, dans les conclusions de son bilan, une telle orientation. Même si la notion de Q.I fait actuellement débat en France (voir par exemple l'article *Gare au Q.I*, publié dans Le journal des psychologues, n° 230 de septembre 2005), la référence à ce type d'indicateur est encore d'usage largement majoritaire dans ces situations.
3. Un autre usage des tests, reposant sur les mêmes épreuves et sur le même type d'indicateur (Q.I), **concerne le diagnostic des élèves dits « surdoués » ou « intellectuellement précoces ».** Si un enfant présente des signes d'un possible « surdouement », les parents (voire les enseignants) peuvent demander au conseiller d'orientation-psychologue

1. Pour une présentation plus complète : Huteau (2006).

de procéder à un examen individuel afin d'établir un diagnostic sur ce plan. Là encore, c'est essentiellement les échelles de Wechsler qui seront utilisées[1]. Mais rappelons que l'indice de Q.I n'est pas à lui seul suffisant pour diagnostiquer une précocité mentale (voir les éléments que nous avons présentés sur ce point dans le chapitre 2 de ce livre).

4. Enfin, mais c'est plus rare, le COP peut être amené à utiliser des **tests d'intelligence logique auprès d'adultes**. C'est le cas, par exemple, d'étudiants en échec à l'Université qui s'interrogent sur les raisons de cet échec ou encore d'adultes en demande de reconversion qui consultent un Centre d'Information et d'Orientation (lieu principal d'activité des COP).

Dans le recrutement et les ressources humaines

Le recours à l'usage des tests afin de sélectionner des personnes pour occuper un emploi donné, est l'une des toutes premières pratiques de la psychologie du travail (Vrignaud et Loarer, 2008). En France, l'idée d'utiliser la psychotechnique pour mieux mettre en relation individus et emplois, est envisagée dès 1904 par Édouard Toulouse qui y voyait un moyen de « classer les individus suivant leurs aptitudes avec une précision bien autre que celle que peuvent fournir des examens superficiels » (cité par Huteau, 2004). Cette idée conduira à la création en 1928 de l'Institut National d'Orientation Professionnelle par Henri Piéron et à l'apparition des premiers services de psychologie dans des grandes entreprises comme Renault (1928) et Citroën (1929). On doit également évoquer les travaux précurseurs de Jean-Marie Lahy, fondateur du Laboratoire Psychotechnique de la STCRP (future RATP) en 1924 et de celui de la Compagnie des chemins de Fer du Nord (puis SNCF) en 1932. Il s'agissait, dans tous les cas, de créer et valider les épreuves psychométriques utiles à une sélection optimale des personnels.

Dans un premier temps, les auteurs ont cherché à évaluer l'ensemble des aptitudes requises dans les différents emplois. Certaines batteries comprenaient près d'une cinquantaine de tests à partir desquels on pouvait constituer des batteries plus restreintes adaptées aux besoins de sélection dans tous les emplois. Les tests les plus pertinents étaient sélectionnés *a priori* (à partir d'une analyse de l'emploi) ou *a posteriori* (à partir des

1. Ce diagnostic peut également être réalisé par un cabinet libéral, le plus souvent avec les mêmes épreuves (de type Q.I).

résultats de personnes déjà en place) (Vrignaud et Loarer, 2008). À partir des années 1960, les batteries se sont simplifiées. La raison principale en est méthodologique : les avancées des méthodes d'analyse factorielle, en particulier l'analyse factorielle confirmatoire, ont abouti à la construction de modèles plus synthétiques (une hiérarchie de facteurs) et plus économiques qui ont conduit à éliminer de nombreuses aptitudes, en fait très redondantes (voir chapitre 1 de cet ouvrage).

➤ La place des tests d'intelligence dans les pratiques de recrutement

Les méthodes de recrutement sont multiples et les recruteurs peuvent s'appuyer sur des techniques variées d'évaluation des personnes. Les études menées dans ce domaine, permettent de saisir cette diversité de techniques et de pratiques. Les tests sont fréquemment utilisés mais l'examen de la littérature internationale montre que leur usage varie fortement selon les pays et selon les praticiens. Ils ne figurent notamment pas dans ce que Cook (1988, cité par Lévy-Leboyer, 2002) appelle le « trio classique » des techniques les plus systématiquement mises en œuvre pour réaliser un recrutement que sont : l'entretien, l'analyse des données biographiques (CV, fiches de renseignement standardisées...), et les références (attestations, recommandations...). Une étude menée en France par Bruchon-Schweizer et Ferrieux (1991) auprès de 102 responsables de recrutement en cabinets et entreprises, est illustrative de ce phénomène. Les résultats en sont présentés dans le tableau 8.1.

Tableau 8.1
Classement des méthodes d'évaluation utilisées en France dans le recrutement selon le % total des services les utilisant (d'après Bruchon-Schweizer et Ferrieux, 1991).

	Fréquence d'emploi		
Techniques utilisées	Systématique	Occasionnel	Total
Entretien(s)	95	4	99
Examen graphologique approfondi	55	38	93
Tests d'aptitudes ou d'intelligence	31	32	63
Tests de personnalité	35	26	61
Mini-situations de travail	7,5	26,5	34
Techniques projectives	12	8,5	20,5
Autres techniques (morphopsychologie, astrologie...)	9	6	15

Les deux techniques auxquelles les services de recrutement ont le plus systématiquement recours sont l'entretien (95 %) et l'examen graphologique approfondi (55 %). Les tests de personnalité viennent ensuite (35 %). Les tests d'intelligence ne sont utilisés systématiquement que par 31 % d'entre eux. D'autres études ont confirmé ces résultats (Vom Hofe et Lévy-Leboyer, 1993, Ballico, 1997, 1999). Ce classement des techniques selon la fréquence de leur usage varie cependant fortement d'un pays à l'autre. Une étude de Dany et Torchy (1994) comparant les pratiques de recrutement dans 12 pays européens, montre ainsi que l'utilisation des tests cognitifs est plus fréquente qu'en France dans une majorité de pays européens, et que la France est seule à avoir un recours si important à la graphologie.

➤ La valeur pronostique des épreuves de recrutement

La question essentielle en matière de recrutement est de savoir dans quelle mesure les évaluations réalisées fournissent des informations sur les performances professionnelles futures du candidat.

Les informations sur les caractéristiques psychologiques des personnes recueillies à l'aide des tests ne prennent de sens qu'en fonction du poste pour lequel on cherche à recruter. Cette mise en relation du profil de personnes avec les exigences d'un emploi peut se faire selon des procédures plus ou moins explicites et validables. Elles peuvent également varier selon la manière dont les tests sont utilisés dans la procédure : par exemple si les tests sont utilisés pour faire une présélection, c'est-à-dire constituer une liste réduite à partir d'un ensemble important de candidats ou si les tests sont utilisés pour classer un petit nombre de candidats préalablement sélectionnés dans un ensemble plus important de candidats.

Une première approche, que l'on peut qualifier de classique par son ancienneté historique, est de considérer que les tests évaluent des caractéristiques générales relativement décontextualisées et, de ce fait, pertinentes pour tous les emplois. Dans le domaine cognitif, on utilisera alors une épreuve de facteur G ou une batterie de tests d'aptitudes et cela quel que soit l'emploi concerné par le recrutement. Cette pratique, qui peut paraître caricaturale reste encore aujourd'hui courante car elle présente un caractère économique et permet de contribuer efficacement à une prise de décision relativement équitable et pertinente, notamment lorsqu'il s'agit de faire un premier tri parmi un grand nombre de candidats. Néanmoins, et bien que la validité prédictive du facteur G, relativement à la réussite professionnelle soit élevée, comme en attestent de nombreuses études, cette méthode ne

saurait suffire pour estimer précisément l'adéquation d'une personne avec un emploi donné.

Une approche alternative, et souvent complémentaire, consiste, à partir d'une analyse précise du travail, à identifier les caractéristiques spécifiquement requises dans l'emploi (capacités ou aptitudes, comportements, "savoir être", etc.) afin de les évaluer de façon ciblée. Cette approche présente l'avantage d'une meilleure proximité entre les épreuves utilisées et les activités menées dans l'emploi mais possède également des limites. D'une part il est difficile de repérer ces éléments car ils ne sont pas directement visibles (seuls les comportements et les performances le sont) et ne peuvent qu'être inférés. Bien que certains prétendent, par des démarches parfois simplistes, y parvenir aisément, ce passage des activités aux « qualités humaines » que ces activités mobilisent demande en réalité la mise en œuvre de démarches relativement complexes, rigoureuses et contrôlées, qui dépassent largement le simple « bon sens »[1]. Elle doit être étayée par des cadres conceptuels solides et validés, afin d'éviter des énoncés flous et donc peu informatifs ou ne présentant pas de cohérence entre eux. D'autre part, il n'existe pas toujours d'épreuves psychométriques permettant d'évaluer de façon valide les caractéristiques identifiées.

Pour savoir avec quelle précision une technique d'évaluation permet de connaître les capacités d'une personne à réussir dans un emploi, on évalue la cohérence statistique entre les résultats d'une épreuve (prise comme prédicteur) et des indicateurs de réussite professionnelle (prises comme critères). Le coefficient de corrélation appelé alors coefficient de prédiction et sa valeur maximum est donc de 1. L'approche pronostique de la validité, quel que soit l'outil considéré, consiste à mettre en relation les résultats du test au moment de la sélection, avec des informations (performances professionnelles par exemple) obtenues en un temps T+1, situé plusieurs mois ou années après le recrutement. Cette approche pose de nombreux problèmes méthodologiques (cf. Vrignaud et Loarer, 2008).

> ***Exemple de difficultés méthodologiques des études de validité pronostique des tests de recrutement***
>
> *Une première difficulté* est que l'échantillon utilisé est dans la plupart des cas constitué d'une partie seulement de l'échantillon de départ. Ce sous-échantillon a en outre été sélectionné en utilisant l'instrument que l'on cherche à valider. Outre le problème de la dépendance entre l'instrument et la procédure, ces données sont censurées c'est-à-dire que les candidats sélectionnés ne sont pas répartis sur toute l'étendue de la distribution mais se situent dans les scores les plus élevés ou du moins proches des seuils jugés optimaux. La variance du sous-échantillon va donc se trouver réduite ce qui aura, entre

> autres conséquences, celle de réduire l'étendue potentielle des corrélations avec les variables critérielles utilisées.
>
> *Une seconde difficulté* méthodologique est le choix du critère retenu pour la validité pronostique qui pose la question de la manière dont se fait l'évaluation de la performance dans l'emploi. Dans certains cas, il est possible de trouver des critères relativement objectifs, comme par exemple le volume de vente pour des épreuves de sélection aux métiers de la vente. Mais cela n'est pas toujours réalisable pour tous les emplois. Bien souvent on va se baser sur une information déclarative, provenant soit de la hiérarchie, soit du candidat lui-même (par exemple sur la satisfaction dans le poste). La fidélité de ces critères n'est pas très élevée, en tout cas, moindre que celle des tests. Or la corrélation entre deux variables est limitée par les valeurs des fidélités de ces variables. (Vrignaud et Loarer, 2008, p. 354.)

De nombreux chercheurs se sont mobilisés pour mesurer les coefficients de prédiction de différentes techniques d'évaluation par rapport à la réussite professionnelle. Des résultats souvent hétérogènes ont été obtenus. Cette hétérogénéité s'explique par la qualité des critères retenus (il est difficile d'échantillonner et de standardiser les indicateurs de réussite professionnelle) ainsi que par les effectifs, souvent limités, des recherches réalisées. Pour limiter ces difficultés, des synthèses (appelées méta-analyses) ont été produites. Il s'agit d'études qui agglomèrent les résultats de nombreuses recherches réalisées sur une même question et comportant des conditions comparables.

Plusieurs méta-analyses ont ainsi pu être réalisées, notamment par Hunter et Hunter, 1984 ; Hunter et Schmidt, 1996 ; Schmidt et Hunter, 1998 ; Robertson et Smith (2001) ; Salgado et al. (2003). Elles fournissent des comparaisons intéressantes et fiables des capacités respectives des différentes techniques d'évaluation à pronostiquer la réussite professionnelle.

Comme le montrent les résultats présentés dans le tableau 8.2 ci-dessous issus d'études qui portent sur plus de 30 000 personnes (d'après Schmidt et Hunter, 1998 et Robertson et Smith, 2001), les mises en situations professionnelles sur des situations sélectionnées présentent généralement les meilleurs coefficients de prédiction (0,54). Cela peut se comprendre par la proximité élevée qu'il y a entre la situation d'évaluation et la situation professionnelle. Le prédicteur est alors très proche du critère. Cette technique est cependant généralement coûteuse et difficile à mettre en œuvre et ne garantit en rien l'adaptation de la personne à long terme, notamment lorsque les caractéristiques du travail évoluent ou lorsque la personne est amenée à changer d'emploi. Les tests d'intelligence générale, outre leur facilité de mise en œuvre, donnent une bonne prédiction de la performance

au travail (0,51). Ils sont supérieurs à l'ensemble des autres techniques utilisables (que se soient les assessment center, les épreuves de personnalité, l'ancienneté au travail ou l'expérience professionnelle). L'entretien présente une validité faible lorsqu'il n'est pas structuré (0,14 à 0,23, Robertson et Smith, 1989), mais cette validité s'améliore si l'entretien est structuré (0,51). On remarquera que les questionnaires de personnalité et d'intérêt présentent des validités plus faibles que les tests d'aptitude cognitive. On notera également la validité nulle de la graphologie. Ce dernier résultat a été souvent démontré (Huteau, 2005) et la persistance de certains recruteurs à utiliser cette méthode non valide ne peut manquer d'étonner. Comme le met en relief l'étude de Bruchon-Schweizer et Ferrieux (1991), ce sont les méthodes les moins valides qui ont tendance, en France, à être les plus utilisées.

Tableau 8.2
Validité prédictive de différentes techniques d'évaluation utilisées en recrutement (selon Schmidt & Hunter, 1998 et Robertson & Schmidt, 2001).

Méthodes d'évaluation	Validité prédictive (corrélation avec la performance au travail)
Échantillons de travail	0,54
Tests d'intelligence (aptitude mentale générale)	0,51
Entretiens structurés	0,51
Évaluations par des pairs	0,49
Tests de connaissances professionnelles	0,48
Tests d'intégrité	0,41
Assessment centers	0,37
Inventaires biographiques	0,35
Tests de personnalité	0,31
Références	0,26
Expérience professionnelle antérieure (nb d'années)	0,18
niveau de scolarité (nb d'années)	0,10
Évaluation des intérêts	0,10
Graphologie	0,02

Étant donné le caractère plus économique en temps de passation et de cotation et la possibilité de passations collectives, on peut considérer que les

tests d'intelligence présentent le meilleur compromis validité/coût. Hunter et Schmidt (1996) ont en outre montré que cette prédiction était générale à l'ensemble des tâches et des professions, ce qui assure la validité à long terme de la prédiction. Une autre méta-analyse menée par Salgado et al. (2003) sur 89 études portant sur des échantillons européens a largement confirmé ces résultats et montré que l'intelligence générale est un bon prédicteur, non seulement de la performance au travail, mais aussi de l'efficacité d'une formation.

Par ailleurs, Hunter et Hunter (1984) et Hunter et Schmidt (1996) ont apporté un élément de validité supplémentaire en montrant (méta-analyse portant sur plus de 400 études) que la validité de la prédiction par les tests d'intelligence augmente avec le degré de complexité de la tâche. Ainsi, de 0,38 pour des tâches de complexité réduite (20 % des emplois étudiés), elle passe à 0,51 pour des tâches de complexité moyenne (63 % des emplois) et à 0,57 pour des tâches de grande complexité (17 % des emplois).

Après avoir comparé les validités des différentes techniques d'évaluation les chercheurs se sont également intéressés à la possibilité de combiner différentes épreuves afin d'améliorer le pronostic global de la performance professionnelle. Plusieurs recherches (par exemple Ree et collaborateurs, 1991, 1994) ont montré que la prise en compte des capacités cognitives spécifiques (aptitudes primaires verbales, numériques, spatiales, mécaniques) ne fournissait pas d'information complémentaire substantielle dans la prédiction des performances professionnelles. Par contre, lorsqu'il s'agit d'épreuves évaluant des dimensions autres que cognitives, des possibilités de gains de validité existent. Ainsi Hunter et Hunter (1984) et Hartigan et Wigdor (1989) ont constaté que les aptitudes psychomotrices augmentaient la validité des tests de facteur G pour les emplois à faible niveau de complexité. De leur côté, Schmidt et Hunter (1998) ont constaté que l'ajout à un test d'intelligence générale d'un test d'intégrité (évaluant le risque de comportements contre productifs tels que voler, se battre, abîmer le matériel...) donne une validité totale de 0,63 (amélioration de 0,12). L'ajout d'un test du caractère consciencieux (l'un des facteurs de personnalité du modèle en 5 facteurs appelé « Big five ») produit également un gain de validité et donne une validité totale de 0,65 (amélioration de 0,14). Ces résultats plaident donc pour une combinaison d'épreuves, sachant que l'essentiel est apporté par l'épreuve d'intelligence générale.

➤ Comment interpréter ces résultats ?

L'interprétation des coefficients de prédiction issus de ces méta-analyses peut être guidée par deux considérations concernant la taille de ces coefficients.

Un coefficient de 0,50 correspond à une fraction de variance (des performances professionnelles) expliquée (par le test) de 25 % (c'est-à-dire r^2). On peut donc considérer que les tests d'intelligence générale « expliquent » environ $\frac{1}{4}$ de la performance professionnelle. C'est plus que ne le font la plupart des autres prédicteurs mais cela peut sembler peu.

Une première remarque est que les méta-analyses sont basées sur un ensemble d'emplois très variés et cela peut contribuer à produire des coefficients de prédiction moins élevés que lorsqu'on travaille sur des emplois très homogènes. Cela est vrai pour l'ensemble des techniques présentant une bonne validité.

Une seconde remarque est que le fait de prendre appui pour étayer une décision de recrutement sur des épreuves présentant ces niveaux de validité améliore très sensiblement la qualité du choix réalisé et réduit fortement le risque d'erreur. Un ensemble important de recherches menées notamment par Rosenthal et ses collègues (Rosenthal et Rubin, 1982 ; Rosenthal et DiMatteo, 2001) dans le domaine médical a montré la pertinence d'une analyse de ces coefficients de prédiction en termes de risques relatifs entre des solutions différentes de comportement. Ainsi, ces auteurs montrent (cité par Rolland, 2004) qu'un coefficient de prédiction de 0,60 entre le respect du traitement d'une maladie et le risque de décès dû à cette maladie équivaut à un risque relatif de 4, c'est-à-dire que le patient qui ne suit pas le traitement a 4 fois plus de chances de mourir que dans le cas contraire. Sur le même principe, on peut considérer que le recours à des tests d'intelligence pour décider d'un recrutement, permet de faire environ 3,3 fois moins d'erreurs que si l'on utilise pour cela une épreuve non valide. Ce ratio est considérable, surtout lorsque l'on considère le coût humain et financier d'une erreur de recrutement.

➤ Tests d'intelligence et respect de la loi contre les discriminations

Une dernière remarque est relative à la notion d'équité dans le processus de sélection. Alors que la lutte contre les discriminations à l'embauche est devenue un enjeu social majeur et une préoccupation essentielle des professionnels du recrutement, le fait de disposer de méthodes valides et équitables est primordial. La loi relative à la lutte contre les discriminations

du 16 novembre 2001 inclut la question des discriminations relatives aux procédures de recrutement et précise que les discriminations visées peuvent être directes ou indirectes. On entend par discrimination indirecte « une discrimination qui se produit lorsqu'une disposition, un critère ou une pratique apparemment neutre, est susceptible d'entraîner un désavantage particulier pour des personnes par rapport à d'autres personnes, à moins que cette disposition, ce critère ou cette pratique apparemment neutre ne soit objectivement justifiée par un objet légitime, et que les moyens de réaliser cet objectif ne soient appropriés et nécessaire » (cité par Gavand, 2006). Les pratiques d'évaluations en vue de recrutement sont bien évidemment susceptibles de produire des discriminations indirectes à l'encontre de certains candidats, en particulier si elles sont affectées de biais avantageant ou désavantageant systématiquement certaines catégories de candidats sur des bases non justifiées par l'objectif de l'évaluation.

La notion de test équitable n'est pas nouvelle. On dit qu'un test est équitable lorsqu'il aboutit à prendre la même décision (par exemple le recrutement) indépendamment des caractéristiques des sujets (par exemple le genre, l'origine sociale, ethnique) (Vrignaud et Loarer, 2008). Le concept d'équité (fairness en anglais) a pour origine un questionnement ancien sur les limites de l'utilisation des tests en général et pour le recrutement en particulier. Il a donné lieu à des développements méthodologiques très sophistiqués pour identifier et réduire les biais pouvant induire des différences de résultats entre les groupes composant la population. Aux États-Unis, plusieurs procès intentés par des victimes de discrimination, ont conduit les utilisateurs ainsi que les éditeurs et les auteurs de tests à procéder à la recherche des biais et à publier dans le manuel des résultats de ces analyses. La publication de ces informations est d'ailleurs explicitement inscrite dans les « recommandations à l'usage des utilisateurs de tests » (voir annexes).

De nombreux travaux ont déjà été réalisés, notamment sur les tests d'intelligence et méritent d'être poursuivis pour parfaire la validité des tests. Sans rentrer ici dans le détail des différents types de biais et des méthodes pour les contrôler (voir le chapitre 2 de cet ouvrage, ainsi que Vrignaud, 2002a) nous retiendrons simplement que de nombreux travaux ont montré la supériorité des méthodes structurées et standardisées pour prévenir ou réduire le risque de biais et qu'il est clair que les tests cognitifs, présentent, de ce point de vue, de nombreux avantages en situation de recrutement par rapport à des situations moins standardisées qui laissent une place plus grande à la subjectivité de l'évaluateur.

Dans les pratiques de conseil, d'accompagnement et d'orientation tout au long de la vie

Depuis les années 1980, les pratiques d'orientation, traditionnellement à destination des élèves, se sont développées et touchent maintenant les adultes et les seniors (Guichard et Huteau, 2006). On parle ainsi d'orientation tout au long de la vie. Dans ces pratiques de conseil, les tests d'intelligence sont utilisés. Bien que les épreuves soient, le plus souvent, identiques à celles utilisés dans les pratiques de recrutement (Laberon, Lagabrielle et Vonthron, 2005), l'objectif du psychologue est, comme nous allons le développer, foncièrement différent. Il ne va pas s'agir en effet d'utiliser ces outils dans un objectif de différenciation, de sélection, mais dans un objectif d'aide à la connaissance de soi, voire d'aide au développement de soi...

La posture même du psychologue va changer dans ce contexte : il va passer de la position de l'expert à celui de conseiller : un conseiller qui ne donne pas obligatoirement de conseils, qui ne se place pas, ou pas exclusivement, en position d'expert, mais plutôt un conseiller qui « tient conseil » avec son consultant (Lhotellier, 2000). Ainsi l'évaluation de type examen psychologique traditionnel (examen psychotechnique) et l'évaluation de type bilan de compétences, approche plus récente, se distinguent sur la place donnée (attribuée) au sujet : dans l'examen la personne est objet de l'évaluation, dans le bilan elle devient partie prenante du processus d'évaluation et d'auto-évaluation (Blanchard, 2002 et 2007).

➤ Qu'est-ce qu'un bilan ?

Il faut comprendre ici cette notion de bilan comme reflétant une approche globale du sujet, dépassant le cadre d'un simple examen psychologique, pour prendre en compte un vaste ensemble de variables, afin de tenter de cerner les différents aspects d'une personne. Aubret et Blanchard, dans leur ouvrage consacré à la pratique du bilan personnalisé (2005) nous en proposent une définition : « la notion de bilan personnalisé renvoie à une approche globale de la personne considérée dans son unité, ses identités, ses rapports à soi, au travail, à autrui, ses valeurs, confrontée à des évènements ou à des situations problèmes impliquant recherche de solutions et prises de décisions » (p. 1).

L'une des formes possibles du bilan personnalisé appliqué à l'adulte est le bilan de compétences.

➤ Le bilan de compétences

Le bilan de compétences est institué en France depuis 1991 – la France semble d'ailleurs en avance à ce niveau comparativement à d'autres pays, comme l'Allemagne par exemple (Eckert et al., 2008) – et est régi par des textes réglementaires. Ainsi, une loi de 1991 instituant le bilan de compétences en défini les objectifs : « les actions permettant de réaliser un bilan de compétence ont pour objet de permettre à des travailleurs d'analyser leurs compétences professionnelles et personnelles ainsi que leurs aptitudes et leurs motivations afin de définir un projet professionnel et, le cas échéant, un projet de formation » (cité par Aubret et Blanchard, 2005, p. 23)

Pour effectuer ce bilan, les professionnels sont, le plus souvent, des psychologues.

➤ Les étapes du bilan de compétence.

D'une durée maximale de 24 heures, le bilan de compétences comporte trois phases :

1. Une phase d'accueil et d'information. On apporte au bénéficiaire une information adaptée à sa demande, on précise cette demande, on vérifie que le bilan correspond aux besoins et attentes de la personne ;
2. Une phase d'investigation. C'est la phase la plus longue au cours de laquelle différentes méthodes et outils (dont les tests) peuvent être utilisés afin d'améliorer 1° la connaissance de soi du bénéficiaire, 2° la connaissance de son environnement professionnel. Le(s) projet(s) personnel(s) commence(nt) à prendre forme ;
3. Une phase de conclusion. Le conseiller et le bénéficiaire hiérarchisent les différents projets, et examinent les conditions de leur réalisation. Un document de synthèse est remis au bénéficiaire.
4. C'est donc dans la deuxième phase du bilan que le psychologue va utiliser des tests.

➤ L'utilisation des tests dans une pratique de bilan

Dans le cadre du bilan, et de la construction d'une « alliance de travail » entre le bénéficiaire du bilan et le psychologue, ce dernier « propose des techniques d'évaluation visant à développer chez lui une meilleure connaissance de

soi, de ses compétences professionnelles et personnelles et une meilleure connaissance du monde professionnel » (Blanchard, 2007, p. 65).

La situation du sujet évalué est donc assez particulière dans le cadre du bilan : il n'est plus un sujet passif mais un sujet actif : « l'individu devient un partenaire, voire un acteur de la gestion de sa carrière » (Guichard et Huteau, 2006, p. 281).

Les psychologues intervenant en bilan utilisent massivement des épreuves destinées à évaluer les aspects conatifs de la personne : motivation, personnalité, intérêts professionnels (Blanchard, Sontag et Leskow, 1999) et également, mais de manière souvent moins systématique, des épreuves cognitives pour évaluer les acquis, aptitudes et compétence. La place donnée aux tests d'intelligence dans une pratique de bilan peut également varier selon les professionnels du bilan. Dans une enquête comparative sur les pratiques d'évaluation en recrutement et en bilan de compétences, nous pouvons relever les constats suivants (Laberon, Lagabrielle et Vonthron, 2005) :

– même si les objectifs des deux situations diffèrent, les méthodes et outils utilisés sont similaires ;
– les tests d'aptitudes sont utilisés par 84 % des conseillers bilans interrogés.

Par contre, aucune information n'est donnée dans cet article sur les tests utilisés. D'après nos propres constats, nous pouvons avancer que les épreuves cognitives les plus utilisées dans le contexte du bilan de compétences sont les batteries factorielles, telles que la NV5 et la NV7, mais également, en fonction des caractéristiques du bénéficiaire (et essentiellement de son niveau de formation), des épreuves de facteur g comme par exemple le D2000 ou encore le R2000.

Dans la formation des adultes

L'usage des tests s'explique ici essentiellement par leur pouvoir prédictif[1] quant à la réussite en formation. Dans le processus de sélection des demandeurs de formation, mais aussi dans le cadre de conseil, il est judicieux d'intégrer des tests d'intelligence logique. C'est le cas, par exemple, pour les formations proposées par l'Association Nationale pour la Formation

1. Nous ne reprendrons pas ici la présentation des études de validité présentées dans la partie « recrutement » (voir plus haut).

Professionnelle des Adultes (A.F.P.A). C'est cet exemple que nous allons maintenant rapidement développer.

➤ Qu'est-ce que l'AFPA ?

Crée en 1945, L'AFPA est l'organisme le plus important en France au niveau de la formation qualifiante.

Elle collabore étroitement avec l'ANPE et propose, aux demandeurs d'emploi mais aussi aux salariés et aux entreprises, différents services : orientation, formations, reclassement... La psychologie à l'AFPA, et tout particulièrement la pratique (et la recherche sur) des tests, est une histoire ancienne mais toujours d'actualité (Fraise, 1991 ; Chartier, 2002).

➤ Les psychologues à l'AFPA

Environ 700 psychologues travaillent à l'AFPA. Ils peuvent exercer directement dans les centres régionaux de l'AFPA mais également dans un établissement spécialisé de l'AFPA, l'Institut National de l'Orientation et de l'Insertion Professionnel (I.N.O.I.P). Dans les centres régionaux, ils interviennent dans les procédures d'évaluations, d'accompagnement, de conseil, ainsi que dans différentes activités liées à l'orientation des adultes (conseil en formation, bilans de compétences...).

Au sein de l'INOIP, ils assurent des missions de recherche et d'étude et développent, par exemple, des méthodes et outils utilisables dans les domaines de l'orientation et de la formation. C'est dans ce cadre qu'ils peuvent être amenés à développer des tests psychologiques.

➤ La pratique des tests à l'AFPA

Différentes formes d'évaluation sont pratiquées à l'AFPA, à différents moments de la formation : en amont, en cours de formation et en fin de formation (Chartier, D., 2002). En ce qui concerne les tests d'intelligence, ils sont principalement utilisés en amont de la formation lors de l'évaluation des capacités d'apprentissage des demandeurs de formation. De par leurs qualités psychométriques, et tout particulièrement de leur validité prédictive : « la méthode des tests a été introduite à des fins de prévisions, conjointement à d'autres modes d'investigation : questionnaires de connaissances générales ou professionnelles, et entretien psychologique » (Fraise, 1991, p. 129).

Pour chaque formation une valeur « seuil » a été définie pour chacune des dimensions évaluées par les tests. L'élaboration de ce seuil est le résultat d'une procédure complexe qui tente de combiner une double contrainte : admettre les candidats qui ont de fortes chances de réussir la formation, écarter ceux qui ont une forte probabilité d'échouer. Dans ce type de décision, Fraise, en 1991, distinguait deux types d'erreurs possibles : 1) ne pas admettre une personne qui réussirait, 2) admettre une personne qui ne réussirait pas (Fraise, 1991). Même si la définition de la valeur seuil, et son utilisation, semble moins rigide aujourd'hui qu'auparavant (Chartier, D., 2002), la même problématique se pose encore actuellement :

> « Ce qui importe le plus, lorsqu'on fixe un seuil, n'est pas le classement des stagiaires qui réussissent, mais le fait qu'on essaie de minimiser le nombre de personnes écartées à tort de la formation qu'ils souhaitent tout en n'envoyant pas en formation des personnes dont la probabilité d'échec est trop élevée » (Chartier, D., 2002, p. 9).

Mais ce seuil n'est pas la seule information prise en compte aujourd'hui par le psychologue de l'AFPA : d'autres éléments (motivation, expérience antérieure...) vont intervenir. C'est à partir de cet ensemble d'informations que le psychologue va prendre sa décision sur l'entrée en formation d'un candidat demandeur de formation. On retrouve ainsi à l'AFPA une volonté de garantir un certain niveau d'objectivité dans les procédures de sélection des candidats

Cet exemple d'utilisation de tests dans une procédure de sélection de candidats à une formation, associée à une certaine souplesse dans la prise de décision, illustre bien quel peut être l'apport de ces épreuves dans un processus de formation.

4. Éditeurs de tests

Nous présentons les coordonnées des principaux éditeurs de tests français (classés par ordre alphabétique).

Éditions Delta Expert, 15, bis rue des Pas Perdus, BP-8338, 95804 Cergy Cedex
www.delta-expert.com

Éditions ECPA et EAP, 25, rue de la plaine, 75980 Paris
Tél : 01 40 09 62 62
www.ecpa.fr

Éditions EUROTEST, 1, impasse de la baleine, 75011 Paris
Tél : 01 48 06 25 75
www.eurotests.com
www.eurotests.eu

Éditions HOGREFE France, 75 avenue Parmentier, 75011 Paris
Tél : 01 40 21 42 08
www.hogrefe.fr

Éditions « Qui plus est », 9, rue du Liban, 75020 Paris
Tél : 01 43 66 61 16
www.editionquiplusest.com

Éditions OPP, 112 rue Réaumur, 75002 Paris
Tél : 01 55 34 92 00
www.opp.eu.com

Éditions SHL, 21 - 23 rue de Madrid, 75008 Paris
Tél : 01 53 04 94 44
Fax : 01 53 04 94 45
www.shl.com/shl/fr/

Fiches pratiques

1. Le test DAT5

Présentation du test

Nom du test : Differential Aptitude Test version 5 (DAT 5)
Auteurs : Bennet, G. K., Seashore, H. G. et Wesman, A. G.
Version : Il s'agit de la 5e version de l'épreuve publiée en 1947 aux États-Unis.
Dates d'édition et de rénovation : La version DAT5 a été diffusée en France en 2002
Éditeur : ECPA.
Format : papier/crayon
Type d'épreuve : batterie factorielle
Type d'items : la DAT5 se compose de huit épreuves indépendantes : Raisonnement Verbal, Raisonnement Abstrait, Raisonnement Numérique, Raisonnement Mécanique, Relations Spatiales, Vitesse de Perception et Précision, Orthographe et Grammaire. Chaque épreuve comporte des items représentatifs de l'aptitude évaluée.
Indicateurs : Le psychologue dispose d'un score pour chaque épreuve ainsi que d'un score composite indicateur de l'efficience scolaire.
Populations visées : adolescents scolarisés (3e à Baccalauréat) et adultes.

Passation

Temps de passation : variable selon les épreuves (de 6 à 20 minutes). Pour une passation de toutes les épreuves il faut prévoir un temps total de 1 h 45 environ.
Modalités de passation : individuelle ou collective
Matériel : réponse sur le cahier de passation ; manuel (81 pages).

Modalités de cotation : rapide, 1 point par bonne réponse.
Étalonnages disponibles : Pour les adolescents en fonction de la formation suivie : classe de 3e, classe de 2e et classe de 1re/Terminale ; Pour les adultes trois niveaux : CAP/BEP, Baccalauréat et étalonnage global.

Informations diverses

Ce test a fait l'objet d'une présentation dans le chapitre 5 de ce livre.

2. Les tests de dominos : D48, D70 et D2000

Présentation du test

Nom des tests : D48 ; D70 ; D2000
Auteur : versions adaptées d'une épreuve anglaise de Anstey de 1943.
Éditeur : ECPA.
Dates d'édition et de rénovation : Première version D48, éditée en 1948. Les noms des versions correspondent aux dates des éditions. La dernière version D2000 a donc été éditée en 2000. Les indications de cette fiche concernent la version D2000.
Format : papier/crayon.
Type d'épreuve : facteur g (intelligence fluide).
Type d'items : quelle que soit la version, D48, D70 ou D2000, il s'agit toujours du même type de tâche : le sujet doit trouver la règle de progression, c'est-à-dire définir la (ou les) relation(s) existant entre les faces des différents dominos qui constituent une certaine suite logique, puis appliquer cette (ou ces) règle(s) de progression afin de déterminer les caractéristiques du domino manquant.
Indicateurs : un score unique.
Populations visées : adulte.

Passation

Temps de passation : 20 minutes (40 items)
Modalités de passation : individuelle ou collective
Matériel : cahier de passation ; feuille de passation ; manuel (30 pages).

Modalités de cotation : rapide, 1 point par bonne réponse.
Étalonnages disponibles : un étalonnage adulte (N = 682), de niveau BEP à Baccalauréat + 5 (sans distinction du niveau) ; un étalonnage concours d'entrée niveau Bac (398 sujets)

Informations diverses

Ce test a fait l'objet d'une présentation dans le chapitre 4 de ce livre
Publications relatives à ce test :
Chartier, P. (à paraître en 2008), Les tests dominos (D70 et D2000) : comment dépasser le constat du seul score total ? Exemples d'analyses des réponses, *Pratiques Psychologiques.*
Dickes, P., et Martin, R. (1998). « Les composantes de l'intelligence générale du D70 ». *Psychologie et Psychométrie*, *19 (1)*, 27-51.
Rémy, L. & Gilles, P-Y. (1999). Stratégies de résolution spatiale et numérique du D70. In M. Huteau & J. Lautrey (Eds), *Approches différentielles en Psychologie*. Rennes : P.U.R.

3. Les tests NNAT

Présentation du test

Nom du test : NNAT (Test d'Aptitude Non Verbal de Nagliéri).
Auteur : J. A. Naglieri.
Éditeur : ECPA.
Dates d'édition et de rénovation : Le test NNAT a été édité aux États-Unis en 1996 et adapté en France en 1998.
Format : papier/crayon.
Type d'épreuve : facteur g (intelligence fluide).
Type d'items : Le test NNAT est décliné en 7 formes, correspondant à 7 niveaux de difficulté : de la forme A (élèves de l'école maternelle) à la forme G (élèves de Lycée). Les tâches sont proches des matrices de Raven : sélectionner l'élément qui vient continuer une série proposée. La spécificité de ce test réside dans la distinction de plusieurs types de raisonnement dans l'épreuve : représentation spatiale, raisonnement analogique et raisonnement

en série. Chaque forme du NNAT comporte 38 items, avec cependant une répartition différente des différents types d'items selon les formes.
Indicateurs : un score total et 3 sous-scores (relatifs aux 3 types de raisonnement).
Populations visées : enfants et adultes.

Passation

Temps de passation : 30 minutes (38 items)
Modalités de passation : individuelle ou collective
Matériel : pour chaque forme : cahier de passation ; feuille de réponse auto-scorable. Le manuel (81 pages) est commun aux 7 formes.
Modalités de cotation : rapide, 1 point par bonne réponse.
Étalonnages disponibles : Il faut signaler ici la possibilité d'interpréter la performance du sujet de deux manières : par un étalonnage selon le niveau scolaire, par un étalonnage selon l'âge.

Informations diverses

Ce test a fait l'objet d'une présentation dans le chapitre 4 de ce livre.

4. Le test R2000 (R85)

Présentation du test

Nom du test : Raisonnement 2000
Noms de l'auteur : P. Rennes
Éditeur : ECPA.
Dates d'édition et de rénovation : La version R2000 est la version éditée en 2000 du test R85 de 1985, issu de l'épreuve de raisonnement de Rennes de 1952.
Format : papier/crayon
Type d'épreuve : facteur g (intelligence fluide) et flexibilité cognitive.
Type d'items : Le test R2000 comporte des items de raisonnement sur des supports variés : verbal, numérique et mixte. La pluralité des supports et des tâches (trouver l'intrus d'une série de mots, suites numériques...) permet

d'évaluer une capacité de flexibilité mentale. Le test comporte 40 items. Ce test est assez difficile.
Indicateurs : un score total.
Populations visées : adultes de niveau de formation Bac + 2 minimum.

Passation

Temps de passation : 20 minutes (40 items)
Modalités de passation : individuelle ou collective
Matériel : cahier de passation ; feuille de réponse ; manuel (31 pages).
Modalités de cotation : rapide, 1 point par bonne réponse.
Étalonnages disponibles : deux étalonnages disponibles :
– un étalonnage global, avec distinction selon le sexe,
– un étalonnage de candidats à un concours (sans autre précision).

Informations diverses

Ce test a fait l'objet d'une présentation dans le chapitre 4 de ce livre.

5. Les tests de Raven : la version SPM

Présentation du test

Nom du test : Standard Progressive Matrices (SPM)
Auteur : John C. Raven
Éditeur : EAP.
Dates d'édition et de rénovation : La première version de l'épreuve date de 1938 (connue sous l'appellation PM38). La dernière version éditée en France date de 1998.
Format : papier/crayon
Type d'épreuve : facteur g (intelligence fluide)
Type d'items : la version SPM comporte 60 items, organisés en cinq séries de 12 items. Les séries sont présentées selon leur niveau de difficulté. Il s'agit toujours du même type de tâche : sélectionner l'élément qui vient continuer une série. Pour pouvoir réussir, le sujet doit trouver la règle de progression,

c'est-à-dire définir les relations existant entre les différents éléments d'une matrice afin de sélectionner l'élément qui vient compléter la série proposée.
Indicateurs : un score unique
Populations visées : adolescents et adultes de niveau de formation jusqu'à bac +2 (au-delà il est préférable d'utiliser la version APM).

Passation

Temps de passation : variable selon l'étalonnage : de 20 à 30 minutes. Possibilité également de passation en temps libre.
Modalités de passation : collective ou individuelle.
Matériel : cahier de passation ; feuille de réponse auto-scorable. Deux manuels : l'un commun avec les autres tests de Raven (manuel section 1 de 96 pages) qui constitue une introduction aux différents tests, l'autre spécifique à cette version SPM (manuel section 3 de 80 pages).
Modalités de cotation : rapide, 1 point par bonne réponse.
Étalonnages disponibles : les étalonnages sont nombreux mais très variés tant au niveau des populations (enfants, scolaires, adultes, de différentes nationalités...) qu'au niveau des conditions de passation (temps limité à 20 ou 30 minutes, temps libre...). Le praticien devra repérer, avant de faire passer l'épreuve, l'étalonnage qui lui semble le plus adapté afin de respecter les conditions de passation de celui-ci (en particulier le temps de passation). Notons qu'il existe un complément d'étalonnage diffusé en 2003.

Informations diverses

Ce test a fait l'objet d'une présentation dans le chapitre 4 de ce livre. Il existe une version plus difficile : les APM.
Principales publications relatives à ce test :
De nombreuses références de publications figurent dans le manuel.
On peut citer également, en langue française :
Raven, J. (2001), Les Progressives Matrices de Raven. Changement et stabilité à travers les cultures et le temps, In M. Huteau, *Les figures de l'intelligence*. Paris : EAP.
Vigneau, F., Douglas, A. B. et Stokes, T. L. (2001), La multidimensionnalité d'un test de facteur g ? Vers une approche expérimentale du test des Matrices de Raven, In A. Flieller, C. Bocéréan, J-L. Kop, E. Thiébaut, A-M. Toniolo et J. Tournois (Eds.), *Questions de psychologie différentielle*. Rennes : PUR.

6. Les tests de Raven : la version APM

Présentation du test

Nom du test : *Advanced Progressive Matrices* (APM)
Auteur : John C. Raven.
Éditeur : EAP.
Dates d'édition et de rénovation : La première version de l'épreuve date de 1943, rénovée en 1947 (connue sous l'appellation PM47). La dernière version a été éditée en France en 1998.
Format : papier/crayon.
Type d'épreuve : test de facteur g (intelligence fluide).
Type d'items : la version APM comporte deux séries d'items dont l'une, le set 1, est destinée à familiariser le sujet avec la situation d'évaluation. La seconde série (set II), qui constitue réellement le test APM, comporte 36 items Les items sont présentés selon leur niveau de difficulté. Il s'agit de tâches analogues à celles de la version SPM (certains items sont communs) : sélectionner l'élément qui vient continuer une série. Pour pouvoir réussir, le sujet doit trouver la règle de progression, c'est-à-dire définir les relations existant entre les différents éléments d'une matrice afin de sélectionner l'élément qui vient compléter la série proposée.
Indicateurs : un score unique
Populations visées : Cette version, plus difficile que la version SPM, concerne les adultes de niveau minimum Bac + 2.

Passation

Temps de passation : 40 minutes ou en temps libre (en fonction de l'étalonnage sélectionné).
Modalités de passation : collective ou individuelle.
Matériel : cahier de passation ; feuille de réponse auto-scorable. Deux manuels : l'un commun avec les autres tests de Raven (manuel section 1 de 96 pages), l'autre spécifique à cette version SPM (manuel section 4 de 126 pages).
Modalités de cotation : rapide, 1 point par bonne réponse.
Étalonnages disponibles : comme pour la version SPM les étalonnages sont nombreux mais assez hétérogènes. Le praticien devra repérer, avant de faire

passer l'épreuve, l'étalonnage qui lui semble le plus adapté afin de définir les conditions de passation (en particulier le temps de passation).

Informations diverses

Ce test a fait l'objet d'une présentation dans le chapitre 4 de ce livre.
Principales publications relatives à ce test :
De nombreuses références de publications figurent dans le manuel.
Raven, J. (2001), Les Progressives Matrices de Raven. Changement et stabilité à travers les cultures et le temps, In M. Huteau, Les figures de l'intelligence. Paris : EAP.

7. Le test Samuel

Présentation du test

Nom du test : SAMUEL
Noms des auteurs : Rozencwajg, P., Corroyer, D**.** et Altman, P.
Éditeur : Delta Expert.
Dates d'édition : Le test SAMUEL a été édité en 1999.
Format : épreuve informatisée
Type d'épreuve : test cognitif : analyse des stratégies de résolution.
Type d'items : Il s'agit d'une adaptation informatisée de la tâche des cubes de Kohs : le sujet doit reproduire sur l'écran de l'ordinateur une configuration modèle à l'aide de plusieurs faces de carrés (unicolores et bicolores) qu'il manipule à l'aide de la souris.
Indicateurs : Le psychologue dispose de deux types d'indicateurs :
– en terme de niveau de réussite,
– en terme de stratégie de résolution.
Concernant le niveau de réussite, les indicateurs sont nombreux : qualité globale de la construction, indice d'anticipation, vitesse d'exécution...
Pour les indices stratégitaires, le psychologue dispose de la stratégie mise en œuvre par le sujet dans chaque item (stratégie globale, analytique ou synthétique) et de la stratégie la plus fréquente sur l'ensemble de l'épreuve (possibilité d'analyse de la variabilité inter et intra-individuelle).
Populations visées : adolescents et adultes.

Passation

Temps de passation : environ 15 minutes.
Modalités de passation : individuelle.
Matériel : logiciel ; manuel (108 pages).
Modalités de cotation : rapide et automatique.
Étalonnages disponibles : par classe d'âge : de 9 ans à l'âge adulte.

Informations diverses

Ce test a fait l'objet d'une présentation dans le chapitre 6 de ce livre.
Principales publications relatives à ce test :
Rozencwajg, P. (2005). *Pour une approche intégrative de l'intelligence. Un siècle après Binet*. Paris : L'Harmattan.
Rozencwajg, P. (2001). « Présentation d'un test cognitif informatisé : SAMUEL », in A. Flieller, C. Bocéréan, J.L. Kop, E. Thiébaut, A.M. Toniolo et J. Tournois (Eds.). *Questions de Psychologie Différentielle* (pages 107-111), Rennes, PUR.

8. Le test TEDE 6

Présentation du test

Nom du test : TEDE6
Auteur : Daniel Pasquier
Éditeur : ECPA
Dates d'édition : Le TEDE6 est la 6e version du test TEDE. Première version en 1989.
Format : papier/crayon
Type d'épreuve : test d'évaluation dynamique saturé en facteur G (intelligence fluide, potentiel d'apprentissage)
Type d'items : Il s'agit d'une adaptation du test des longueurs de Faverge (1955). L'épreuve comprend 12 items d'apprentissage et 18 items de test. Chaque item comprend un double aspect : logico-mathématique d'une part (le sujet doit composer une égalité de longueurs de segments de droite selon la structure additive de type A=B+C) et visuo-projectif d'autre part (le

sujet doit montrer que l'égalité est vraie par superposition des segments en opérant les projections nécessaires).
Populations visées : populations adultes, apprentis ou candidats apprentis sachant lire le français.

Passation

Temps de passation : 2 heures (apprentissage : 1 heure + pause de 15 minutes + test : 45 minutes).
Modalités de passation : collective ou individuelle.
Matériel : Deux livrets (le livret d'apprentissage et le livret de test), un dossier d'instructions et un logiciel de correction.
Correction : Le temps de correction est d'environ 5 minutes à l'aide du logiciel de correction. Ce programme donne accès à différents traitements : positionnement au regard de l'étalonnage adapté, calcul de différents « profils » du sujet.
Étalonnages disponibles :
2 étalonnages : population d'adultes et population d'apprentis.
Des étalonnages spécifiques sont en outre proposés :
– pour la population adulte selon 4 niveaux de qualification :
groupe 1 : les hommes et femmes de niveau II et ID, les hommes de niveau IV, tous âges confondus.
groupe 2 : les hommes juniors (16-29 ans) de niveau V et les femmes juniors (16-29 ans) de niveau IV
groupe 3 : les femmes de niveau V, les hommes seniors (29-58 ans) de niveau V et les femmes seniors (29-58 ans) de niveau IV
groupe 4 : les hommes et femmes de niveau VI, tous âges confondus.
– pour la population des apprentis et candidats apprentis, selon 4 niveaux de qualification des diplômes préparés : CAP, BEP, BP, Bac

Informations diverses

Ce test a fait l'objet d'une présentation dans le chapitre 7 de ce livre.
Principales publications relatives à ce test :
Pasquier D. (1994). Le test d'évaluation dynamique de l'éducabilité (T.E.D.E.), in Huteau M. (Ed.). *Les techniques d'évaluation des personnes.* Issy-les-Moulineaux : EAP.

Pasquier, D. (1995). Le T.E.D.E. *Pédagogies de Médiations, Documents du CRU/SE.* Poitiers : CUFEP.
Pasquier D., Estebe I., Jaigu J. (2001). « Prévoir la réussite chez de jeunes apprentis : approche exploratoire », *Pratiques Psychologiques,* 1, 99-110.
Pasquier, D. (2005). *Manuel d'utilisation et d'interprétation du TEDE 4*. Paris : ECPA

9. Les tests de WECHSLER : le WISC-III

Présentation du test

Nom du test : Wechsler Intelligence Scale for Children version III (WISC-III)
Auteur : David Wechsler.
Éditeur : ECPA.
Dates d'édition et de rénovation : 3e version de l'épreuve de Wechsler pour enfant. Version adaptée en France en 1996.
Format : variable selon les subtests : questionnement du psychologue et tests de performance.
Type d'épreuve : échelle composite d'intelligence pour enfant.
Type d'items : la structure du WISC-III comporte deux échelles, une échelle verbale et une échelle de performance. Chaque échelle est composée de différents subtests qui composent des situations très variées d'évaluation (approche globale de l'intelligence). Pour l'échelle verbale (13 subtests) l'enfant doit répondre oralement à des questions posées par le psychologue : trouver la relation entre deux notions, connaissance de son environnement, petits problèmes arithmétiques... Pour l'échelle de performance (7 subtests) l'enfant doit réaliser différentes tâches : constructions à l'aide de cubes, arrangements d'images en ordre chronologique, assemblages d'éléments de type puzzle...
Dans chaque subtest les items sont présentés selon leur niveau de difficulté.
Indicateurs : comme toutes les échelles de Wechsler, les performances sont exprimées sous forme de QI : un QIT (ou QI Total) et un QI pour chaque échelle (QIV et QIP). Le psychologue dispose également de scores standardisés pour chaque subtest (analyse du profil des résultats).
Pour cette version WISC-III se rajoute la possibilité de calculer 3 indices factoriels : Indice de compréhension verbale (ICV), indice d'organisation perceptive (IOP) et indice vitesse de traitement (IVT).

Populations visées : enfants et adolescents âgés de 6 à 16 ans.

Passation

Temps de passation : 1 h 15 à 1 h 45 (variable selon le niveau de réussite de l'enfant : règles d'arrêt).
Modalités de passation : individuelle.
Matériel : une mallette regroupe l'ensemble du matériel nécessaire dont un manuel de 294 pages.
Modalités de cotation : la cotation est assez complexe. Les indications du manuel seront précieuses. Certaines réponses sont classiquement cotées 0 ou 1 point tandis que pour d'autres (certains subtests de l'échelle verbale) la cotation est plus fine en distinguant les bonnes réponses à 1 point des réponses de qualité supérieures cotées 2 points
Le psychologue calcule une note pour chaque subtest, puis combine ces notes pour obtenir les QI et les indices factoriels. Attention : 10 subtests sont obligatoires pour pouvoir calculer un Q.I
Étalonnages disponibles : étalonnages très précis, par classe d'âge de 4 mois, de type Q.I pour les 3 indicateurs QIT, QIV et QIP comme pour les 3 indices factoriels. Étalonnages pour chaque subtest (score de 1 à 19).

Informations diverses

Ce test a fait l'objet d'une présentation dans le chapitre 3 de ce livre.
Une version plus récente (WISC-IV) est diffusée depuis 2005 en France.
Principales publications relatives à ce test :
Arbisio, C. (2003). *Le bilan psychologique avec l'enfant. Approche clinique du WISC-III.* Paris : Dunod.
Grégoire, J. (2000a). *L'examen clinique de l'intelligence de l'enfant.* Sprimont : Mardaga.

10. Les tests de WECHSLER : le WISC-IV

Présentation du test

Nom des tests : Wechsler Intelligence Scale for Children version IV (WISC-IV)
Auteurs : David Wechsler.
Éditeur : ECPA.
Dates d'édition et de rénovation : 4e version de l'épreuve de Wechsler pour enfant. Version adaptée en France en 2005.
Format : variable selon les subtests : questionnement du psychologue, papier/crayon et tests de performance.
Type d'épreuve : échelle d'intelligence pour enfant.
Type d'items : la structure du WISC-IV est sensiblement différente de celles des anciennes versions du Wechsler pour enfant. En effet disparaissent ici les deux échelles classiques, l'échelle verbale et l'échelle de performance, au profit de quatre indices factoriels : Indice de Compréhension Verbale (ICV), Indices de Raisonnement Perceptif (IRP), Indice de Mémoire de Travail (IMT) et Indice de Vitesse de Traitement (IVT). Seul l'indicateur QIT est conservé. Il s'agit donc plus d'une réelle transformation du WISC que d'une simple rénovation.
Chaque indice est composé de différents subtests qui sont proches des subtests de l'ancienne version WISC-III ou de la version pour adulte WAIS-III. Ces situations d'évaluation restent assez variées : trouver la relation entre deux notions, compréhension de situations de la vie courante, devinettes, petits problèmes arithmétiques, constructions à l'aide de cubes, matrices analogiques... Dans chaque subtest les items sont présentés selon leur niveau de difficulté.
Au total le WISC-IV comporte 15 subtests, certains d'entre eux étant optionnels.
Indicateurs : par rapport aux versions précédentes seul l'indicateur QIT (quotient intellectuel total) est conservé. Le psychologue dispose de 4 indicateurs relatifs aux indices ICV, IRP, IMT et IVT (exprimés dans la même métrique que le QI) ainsi que des indicateurs normalisés pour chaque subtest (analyse du profil des résultats).
Populations visées : enfants et adolescents âgés de 6 à 16 ans 1/2.

Passation

Temps de passation : 1 h 15 à 1 h 45 (variable selon le niveau de réussite de l'enfant : règles d'arrêt).
Modalités de passation : individuelle.
Matériel : une mallette regroupe l'ensemble du matériel nécessaire. Deux manuels accompagnent l'épreuve, l'un destiné à la passation et la cotation (273 pages), l'autre centré sur les qualités psychométriques de l'épreuve et les éléments d'interprétation des scores (123 pages).
Modalités de cotation : la cotation est assez complexe. Les indications du manuel seront précieuses. Certaines réponses sont classiquement cotées 0 ou 1 point tandis que pour d'autres (certains subtests de l'échelle verbale) la cotation est plus fine en distinguant les bonnes réponses à 1 point des réponses de qualité supérieures cotées 2 points.
Le psychologue calcule une note pour chaque subtest, puis combine ces notes pour obtenir les QI et les indices.
Il existe un Cd-rom d'aide à la cotation.
Étalonnages disponibles : étalonnages très précis, par classe d'âge de 4 mois, dans une métrique de type Q.I (m=100 et Écart type de 15) pour le QIT et les quatre indices (ICV, IRP, IMT et IVT). Étalonnage pour les notes aux subtests

Informations diverses

Ce test a fait l'objet d'une présentation dans le chapitre 3 de ce livre.
Principales publications relatives à ce test :
Grégoire, J. (2006). *L'examen clinique de l'intelligence de l'enfant. Fondements et pratique du WISC-IV*. Sprimont : Mardaga.
Rozencwajg, P. (2006). Quelques réflexions sur l'évaluation de l'intelligence générale : un retour à Binet, *Pratiques Psychologiques,* 12 (3), 395-410

11. Les tests de WECHSLER : la WAIS-III

Présentation du test

Nom du test : Wechsler Adult Intelligence Scale version III (WAIS-III)

Auteur : David Wechsler.
Éditeur : ECPA.
Dates d'édition et de rénovation : 3e version de l'épreuve de Wechsler pour adultes. Version adaptée en France en 2000.
Format :
Variable selon les subtests : questionnement du psychologue, papier/crayon et tests de performance.
Type d'épreuve : échelle d'intelligence pour adulte.
Type(s) d'items : la structure de la WAIS-III est comparable à celle du WISC-III : une échelle totale (QIT) et deux sous échelles, verbale (QIV) et performance (QIP). On retrouve également la possibilité de calculer des indices factoriels, ici au nombre de quatre : Compréhension Verbale (I.C.V), Organisation Perceptive (I.O.P), Mémoire de Travail (MT) et Vitesse de Traitement (IVT).
Chaque indicateur (Q.I ou indice factoriel), prend appui sur différents subtests. Les situations d'évaluation sont variées : trouver la définition d'un mot, la similitude entre deux notions, compréhension de situations de la vie courante, petits problèmes arithmétiques, constructions à l'aide de cubes, matrices analogiques... Au total, la version WAIS-III comporte 14 subtests. Dans chaque subtest les items sont présentés selon leur niveau de difficulté.
Indicateurs :
Le psychologue dispose d'un total de 7 indicateurs : les 3 indicateurs de type Q.I (QIT, QIV et QIP) et les 4 indices factoriels (ICV, IOP, IMT et IVT). Il dispose également d'indicateurs plus spécifiques concernant le niveau de réussite dans chaque subtest (analyse du profil des résultats).
Populations visées : adolescents et adultes âgés de 16 à 89 ans.

Passation

Temps de passation : 1 h 15 à 1 h 45 (variable en fonction du niveau de réussite : règles d'arrêt).
Modalités de passation : individuelle.
Matériel : une mallette regroupe l'ensemble du matériel nécessaire. Le manuel est composé de 357 pages.
Modalités de cotation : comme les autres échelles de Wechsler la cotation est assez complexe et le psychologue non expérimenté suivra scrupuleusement les indications (nombreuses) du manuel. On retrouve dans certaines subtests la cotation en trois niveaux : 0, 1 ou 2 points.

Le psychologue calcule une note pour chaque subtest, puis combine ces notes pour obtenir les QI et les indices.
Étalonnages disponibles : étalonnages très précis, par classe d'âge, dans une métrique de type Q.I (m=100 et Écart type de 15) pour les QI et les quatre indices. Étalonnage pour les notes aux subtests.

Informations diverses

Ce test a fait l'objet d'une présentation dans le chapitre 3 de ce livre.
Principales publications relatives à ce test :
Castro, D. (2006). *Pratique de l'examen psychologique en clinique adulte*. Paris : Dunod
Grégoire, J. (2004). *L'examen clinique de l'intelligence de l'adulte.* Sprimont : Mardaga.

12. Le test NV5-R

Présentation du test

Nom du test : NV5-R
Auteurs : Robert Simonet (1987) puis Thiébaut, E et Bidan-Fortier (2003).
Éditeur : EAP.
Dates d'édition et de rénovation : version rénovée de l'épreuve NV5 publiée en 1987, la version NV5-R a été éditée en 2003.
Format : papier/crayon.
Type d'épreuve : batterie factorielle.
Type d'items : la NV5-R est composée de plusieurs tests indépendants. Plus précisément elle comporte 9 épreuves : Raisonnement général, Raisonnement inductif, Raisonnement spatial, Raisonnement pratique/technique, Compréhension verbale, Vocabulaire, Orthographe, Calcul et Attention. Chaque épreuve comporte des items représentatifs de l'aptitude évaluée. L'épreuve de Raisonnement général est un peu atypique ici car elle est composée de différents types d'items.
Indicateurs : Le psychologue dispose d'un score pour chaque aptitude. À partir de ces scores il peut procéder à deux types d'analyse en déterminant : 1° un profil d'aptitude qui comporte 4 scores (aptitude générale, verbale,

spatiale et numérique), 2° un profil cognitif, qui repose sur le modèle théorique du radex, avec distinction de trois niveaux de raisonnement (général, intermédiaire, spécifique).
Populations visées : adultes de niveau minimum Baccalauréat (utiliser la version NV7 pour des niveaux plus faibles).

Passation

Temps de passation : variable selon les épreuves. Pour une passation de toutes les épreuves il faut prévoir un temps total de 1 h 45 à 2 heures environ.
Modalités de passation : individuelle ou collective.
Matériel : cahier de passation ; feuille de réponse auto-scorable ; manuel (108 pages).
Modalités de cotation : rapide, 1 point par bonne réponse.
Étalonnages disponibles : un étalonnage hétérogène avec séparation par niveau d'étude.

Informations diverses

Ce test a fait l'objet d'une présentation dans le chapitre 5 de ce livre.
Publications relatives à ce test :
Thiébaut, E. et Richoux, V. (2005), Éléments de validité prédictive des scores à la batterie d'aptitudes cognitives NV5-R, *Pratiques Psychologiques,* 11, 404-416.

13. Le test NV7

Présentation du test

Nom du test : NV7.
Auteurs : Il s'agit d'une élaboration d'un collectif (Bernaud, Priou et Simonet) à partir de la sélection de tests existants.
Éditeur : EAP.
Dates d'édition : version éditée en 1993.
Format : papier/crayon.
Type d'épreuve : batterie factorielle.

Type d'items : la NV7 est composée de dix épreuves indépendantes : Raisonnement déductif, Raisonnement inductif, Raisonnement analogique, Raisonnement pratique/technique, Spatial, Problèmes, Opérations, Attention, Orthographe et Compréhension verbale. Chaque épreuve comporte des items représentatifs de l'aptitude évaluée.
Indicateurs : Le psychologue dispose d'un score pour chaque épreuve ainsi que de deux scores composites : Efficience Intellectuelle Générale (EIG) et Efficience Scolaire (ES). Il dispose également d'indicateurs concernant la rapidité des réponses et leur précision.
Populations visées : adolescents et adultes de bas niveau de qualification (inférieur au Baccalauréat).

Passation

Temps de passation : variable selon les épreuves. Pour une passation de toutes les épreuves il faut prévoir un temps total de 1 h 45 environ.
Modalités de passation : individuelle ou collective.
Matériel : cahier de passation ; feuille de réponse auto-scorable ; manuel (64 pages).
Modalités de cotation : rapide, 1 point par bonne réponse.
Étalonnages disponibles : jeunes peu qualifiés ; adultes faiblement qualifiés ; jeunes apprentis.

Informations diverses

Ce test a fait l'objet d'une présentation dans le chapitre 5 de ce livre.

Annexes

1. Code de déontologie des psychologues praticiens[1]

Préambule

Le présent Code de Déontologie est destiné à servir de règle professionnelle aux hommes et aux femmes qui ont le titre de psychologue, quels que soient leur mode d'exercice et leur cadre professionnel, y compris leurs activités d'enseignement et de recherche.

Sa finalité est avant tout de protéger le public et les psychologues contre les mésusages de la psychologie et contre l'usage de méthodes et techniques se réclamant abusivement de la psychologie.

Les organisations professionnelles signataires du présent Code s'emploient à le faire connaître et respecter. Elles apportent, dans cette perspective, soutien et assistance à leurs membres. L'adhésion des psychologues à ces organisations implique leur engagement à respecter les dispositions du Code.

Titre I. Principes généraux

La complexité des situations psychologiques s'oppose à la simple application systématique de règles pratiques. Le respect des règles du présent Code de Déontologie repose sur une réflexion éthique et une capacité de discernement, dans l'observance des grands principes suivants :

1. Code signé par l'Association des Enseignants de Psychologie des Universités (AEPU), l'Association Nationale des Organisations de Psychologues (ANOP), la Société Française de Psychologie (SFP) le 22 mars 1996.

➤ 1. Respect des droits de la personne

Le psychologue réfère son exercice aux principes édictés par les législations nationale, européenne et internationale sur le respect des droits fondamentaux des personnes, et spécialement de leur dignité, de leur liberté et de leur protection. Il n'intervient qu'avec le consentement libre et éclairé des personnes concernées. Réciproquement, toute personne doit pouvoir s'adresser directement et librement à un psychologue. Le psychologue préserve la vie privée des personnes en garantissant le respect du secret professionnel, y compris entre collègues. Il respecte le principe fondamental que nul n'est tenu de révéler quoi que ce soit sur lui-même.

➤ 2. Compétence

Le psychologue tient ses compétences de connaissances théoriques régulièrement mises à jour, d'une formation continue et d'une formation à discerner son implication personnelle dans la compréhension d'autrui. Chaque psychologue est garant de ses qualifications particulières et définit ses limites propres, compte tenu de sa formation et de son expérience. Il refuse toute intervention lorsqu'il sait ne pas avoir les compétences requises.

➤ 3. Responsabilité

Outre les responsabilités définies par la loi commune, le psychologue a une responsabilité professionnelle. Il s'attache à ce que ses interventions se conforment aux règles du présent Code. Dans le cadre de ses compétences professionnelles, le psychologue décide du choix et de l'application des méthodes et techniques psychologiques qu'il conçoit et met en œuvre. Il répond donc personnellement de ses choix et des conséquences directes de ses actions et avis professionnels.

➤ 4. Probité

Le psychologue a un devoir de probité dans toutes ses relations professionnelles. Ce devoir fonde l'observance des règles déontologiques et son effort continu pour affiner ses interventions, préciser ses méthodes et définir ses buts.

➤ 5. Qualité scientifique

Les modes d'intervention choisis par le psychologue doivent pouvoir faire l'objet d'une explicitation raisonnée de leurs fondements théoriques et de leur construction. Toute évaluation ou tout résultat doit pouvoir faire l'objet d'un débat contradictoire des professionnels entre eux.

➤ 6. Respect du but assigné

Les dispositifs méthodologiques mis en place par le psychologue répondent aux motifs de ses interventions, et à eux seulement. Tout en construisant son intervention dans le respect du but assigné, le psychologue doit donc prendre en considération les utilisations possibles qui peuvent éventuellement en être faites par des tiers.

➤ 7. Indépendance professionnelle

Le psychologue ne peut aliéner l'indépendance nécessaire à l'exercice de sa profession sous quelque forme que ce soit.

➤ Clause de conscience

Dans toutes les circonstances où le psychologue estime ne pas pouvoir respecter ces principes, il est en droit de faire jouer la clause de conscience.

Titre II. L'exercice professionnel

➤ Chapitre 1. Le titre de psychologue et la définition de la profession

Article 1
L'usage du titre de psychologue est défini par la loi n° 85-772 du 25 juillet 1985 publiée au J.O. du 26 juillet 1985. Sont psychologues les personnes qui remplissent les conditions de qualification requises dans cette loi. Toute forme d'usurpation du titre est passible de poursuites.

Article 2
L'exercice professionnel de la psychologie requiert le titre et le statut de psychologue.

Article 3
La mission fondamentale du psychologue est de faire reconnaître et respecter la personne dans sa dimension psychique. Son activité porte sur la composante psychique des individus, considérés isolément ou collectivement.

Article 4
Le psychologue peut exercer différentes fonctions à titre libéral, salarié ou d'agent public. Il peut remplir différentes missions, qu'il distingue et fait distinguer, comme le conseil, l'enseignement de la psychologie, l'évaluation, l'expertise, la formation, la psychothérapie, la recherche, etc. Ces missions peuvent s'exercer dans divers secteurs professionnels.

➤ Chapitre 2. Les conditions de l'exercice de la profession

Article 5
Le psychologue exerce dans les domaines liés à sa qualification, laquelle s'apprécie notamment par sa formation universitaire fondamentale et appliquée de haut niveau en psychologie, par des formations spécifiques, par son expérience pratique et ses travaux de recherche. Il détermine l'indication et procède à la réalisation d'actes qui relèvent de sa compétence.

Article 6
Le psychologue fait respecter la spécificité de son exercice et son autonomie technique. Il respecte celles des autres professionnels.

Article 7
Le psychologue accepte les missions qu'il estime compatibles avec ses compétences, sa technique, ses fonctions, et qui ne contreviennent ni aux dispositions du présent Code, ni aux dispositions légales en vigueur.

Article 8
Le fait pour un psychologue d'être lié dans son exercice professionnel par un contrat ou un statut à toute entreprise privée ou tout organisme public, ne modifie pas ses devoirs professionnels, et en particulier ses obligations concernant le secret professionnel et l'indépendance du choix de ses méthodes et de ses décisions. Il fait état du Code de Déontologie dans l'établissement de ses contrats et s'y réfère dans ses liens professionnels.

Article 9
Avant toute intervention, le psychologue s'assure du consentement de ceux qui le consultent ou participent à une évaluation, une recherche ou une expertise. Il les informe des modalités, des objectifs et des limites de son intervention. Les avis du psychologue peuvent concerner des dossiers ou des situations qui lui sont rapportées. Mais son évaluation ne peut porter que sur des personnes ou des situations qu'il a pu examiner lui-même. Dans toutes les situations d'évaluation, quel que soit le demandeur, le psychologue rappelle aux personnes concernées leur droit à demander une contre-évaluation. Dans les situations de recherche, il les informe de leur droit à s'en retirer à tout moment. Dans les situations d'expertise judiciaire, le psychologue traite de façon équitable avec chacune des parties et sait que sa mission a pour but d'éclairer la justice sur la question qui lui est posée et non d'apporter des preuves.

Article 10
Le psychologue peut recevoir, à leur demande, des mineurs ou des majeurs protégés par la loi. Son intervention auprès d'eux tient compte de leur statut, de leur situation et des dispositions légales en vigueur. Lorsque la consultation pour des mineurs ou des majeurs protégés par la loi est demandée par un tiers, le psychologue requiert leur consentement éclairé, ainsi que celui des détenteurs de l'autorité parentale ou de la tutelle.

Article 11
Le psychologue n'use pas de sa position à des fins personnelles, de prosélytisme ou d'aliénation d'autrui. Il ne répond pas à la demande d'un tiers qui recherche un avantage illicite ou immoral, ou qui fait acte d'autorité abusive dans le recours à ses services. Le psychologue n'engage pas d'évaluation ou de traitement impliquant des personnes auxquelles il serait déjà personnellement lié.

Article 12
Le psychologue est seul responsable de ses conclusions. Il fait état des méthodes et outils sur lesquels il les fonde, et il les présente de façon adaptée à ses différents interlocuteurs, de manière à préserver le secret professionnel. Les intéressés ont le droit d'obtenir un compte rendu compréhensible des évaluations les concernant, quels qu'en soient les destinataires. Lorsque ces conclusions sont présentées à des tiers, elles ne répondent qu'à la question posée et ne comportent les éléments d'ordre psychologique qui les fondent que si nécessaire.

Article 13
Le psychologue ne peut se prévaloir de sa fonction pour cautionner un acte illégal, et son titre ne le dispense pas des obligations de la loi commune. Conformément aux dispositions de la loi pénale en matière de non-assistance à personne en danger, il lui est donc fait obligation de signaler aux autorités judiciaires chargées de l'application de la Loi toute situation qu'il sait mettre en danger l'intégrité des personnes. Dans le cas particulier où ce sont des informations à caractère confidentiel qui lui indiquent des situations susceptibles de porter atteinte à l'intégrité psychique ou physique de la personne qui le consulte ou à celle d'un tiers, le psychologue évalue en conscience la conduite à tenir, en tenant compte des prescriptions légales en matière de secret professionnel et d'assistance à personne en danger. Le psychologue peut éclairer sa décision en prenant conseil auprès de collègues expérimentés.

Article 14
Les documents émanant d'un psychologue (attestation, bilan, certificat, courrier, rapport, etc.) portent son nom, l'identification de sa fonction ainsi que ses coordonnées professionnelles, sa signature et la mention précise du destinataire. Le psychologue n'accepte pas que d'autres que lui-même modifient, signent ou annulent les documents relevant de son activité professionnelle. Il n'accepte pas que ses comptes rendus soient transmis sans son accord explicite, et il fait respecter la confidentialité de son courrier.

Article 15
Le psychologue dispose sur le lieu de son exercice professionnel d'une installation convenable, de locaux adéquats pour permettre le respect du secret professionnel, et de moyens techniques suffisants en rapport avec la nature de ses actes professionnels et des personnes qui le consultent.

Article 16
Dans le cas où le psychologue est empêché de poursuivre son intervention, il prend les mesures appropriées pour que la continuité de son action professionnelle soit assurée par un collègue avec l'accord des personnes concernées, et sous réserve que cette nouvelle intervention soit fondée et déontologiquement possible.

➤ Chapitre 3 : Les modalités techniques de l'exercice professionnel

Article 17
La pratique du psychologue ne se réduit pas aux méthodes et aux techniques qu'il met en œuvre. Elle est indissociable d'une appréciation critique et d'une mise en perspective théorique de ces techniques.

Article 18
Les techniques utilisées par le psychologue pour l'évaluation, à des fins directes de diagnostic, d'orientation ou de sélection, doivent avoir été scientifiquement validées.

Article 19
Le psychologue est averti du caractère relatif de ses évaluations et interprétations. Il ne tire pas de conclusions réductrices ou définitives sur les aptitudes ou la personnalité des individus, notamment lorsque ces conclusions peuvent avoir une influence directe sur leur existence.

Article 20
Le psychologue connaît les dispositions légales et réglementaires issues de la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. En conséquence, il recueille, traite, classe, archive et conserve les informations et données afférentes à son activité selon les dispositions en vigueur. Lorsque ces données sont utilisées à des fins d'enseignement, de recherche, de publication, ou de communication, elles sont impérativement traitées dans le respect absolu de l'anonymat, par la suppression de tout élément permettant l'identification directe ou indirecte des personnes concernées, ceci toujours en conformité avec les dispositions légales concernant les informations nominatives.

➤ Chapitre 4. Les devoirs du psychologue envers ses collègues

Article 21
Le psychologue soutient ses collègues dans l'exercice de leur profession et dans l'application et la défense du présent Code. Il répond favorablement à leurs demandes de conseil et les aide dans les situations difficiles, notamment en contribuant à la résolution des problèmes déontologiques.

Article 22
Le psychologue respecte les conceptions et les pratiques de ses collègues pour autant qu'elles ne contreviennent pas aux principes généraux du présent Code ; ceci n'exclut pas la critique fondée.

Article 23
Le psychologue ne concurrence pas abusivement ses collègues et fait appel à eux s'il estime qu'ils sont plus à même que lui de répondre à une demande.

Article 24
Lorsque le psychologue remplit une mission d'audit ou d'expertise vis-à-vis de collègues ou d'institutions, il le fait dans le respect des exigences de sa déontologie.

➤ Chapitre 5. Le psychologue et la diffusion de la psychologie

Article 25
Le psychologue a une responsabilité dans la diffusion de la psychologie auprès du public et des médias. Il fait de la psychologie et de ses applications une présentation en accord avec les règles déontologiques de la profession. Il use de son droit de rectification pour contribuer au sérieux des informations communiquées au public.

Article 26
Le psychologue n'entre pas dans le détail des méthodes et techniques psychologiques qu'il présente au public, et il l'informe des dangers potentiels d'une utilisation incontrôlée de ces techniques.

Titre III. La formation du psychologue

➤ Chapitre 1. Les principes de la formation

Article 27
L'enseignement de la psychologie à destination des futurs psychologues respecte les règles déontologiques du présent Code. En conséquence, les institutions de formation :

- diffusent le Code de Déontologie des Psychologues aux étudiants dès le début des études ;
- s'assurent de l'existence de conditions permettant que se développe la réflexion sur les questions d'éthique liées aux différentes pratiques : enseignement et formation, pratique professionnelle, recherche.

Article 28
L'enseignement présente les différents champs d'étude de la psychologie, ainsi que la pluralité des cadres théoriques, des méthodes et des pratiques, dans un souci de mise en perspective et de confrontation critique. Il bannit nécessairement l'endoctrinement et le sectarisme.

Article 29
L'enseignement de la psychologie fait une place aux disciplines qui contribuent à la connaissance de l'homme et au respect de ses droits, afin de préparer les étudiants à aborder les questions liées à leur futur exercice dans le respect des connaissances disponibles et des valeurs éthiques.

➤ Chapitre 2. Conception de la formation

Article 30
Le psychologue enseignant la psychologie ne participe pas à des formations n'offrant pas de garanties sur le sérieux des finalités et des moyens. Les enseignements de psychologie destinés à la formation continue des psychologues ne peuvent concerner que des personnes ayant le titre de psychologue. Les enseignements de psychologie destinés à la formation de professionnels non-psychologues observent les mêmes règles déontologiques que celles énoncées aux Articles 27, 28 et 32 du présent Code.

Article 31
Le psychologue enseignant la psychologie veille à ce que ses pratiques, de même que les exigences universitaires (mémoires de recherche, stages professionnels, recrutement de sujets, etc.), soient compatibles avec la déontologie professionnelle. Il traite les informations concernant les étudiants, acquises à l'occasion des activités d'enseignement, de formation ou de stage, dans le respect des Articles du Code concernant les personnes.

Article 32
Il est enseigné aux étudiants que les procédures psychologiques concernant l'évaluation des individus et des groupes requièrent la plus grande rigueur scientifique et éthique dans leur maniement (prudence, vérification) et leur utilisation (secret professionnel et devoir de réserve), et que les présentations de cas se font dans le respect de la liberté de consentir ou de refuser, de la dignité et du bien-être des personnes présentées.

Article 33
Les psychologues qui encadrent les stages, à l'Université et sur le terrain, veillent à ce que les stagiaires appliquent les dispositions du Code, notamment celles qui portent sur la confidentialité, le secret professionnel, le consentement éclairé. Ils s'opposent à ce que les stagiaires soient employés comme des professionnels non rémunérés. Ils ont pour mission de former professionnellement les étudiants, et non d'intervenir sur leur personnalité.

Article 34
Conformément aux dispositions légales, le psychologue enseignant la psychologie n'accepte aucune rémunération de la part d'une personne qui a droit à ses services au titre de sa fonction universitaire. Il n'exige pas des étudiants qu'ils suivent des formations extra-universitaires payantes ou non, pour l'obtention de leur diplôme. Il ne tient pas les étudiants pour des patients ou des clients. Il n'exige pas leur participation gratuite ou non, à ses autres activités, lorsqu'elles ne font pas explicitement partie du programme de formation dans lequel sont engagés les étudiants.

Article 35
La validation des connaissances acquises au cours de la formation initiale se fait selon des modalités officielles. Elle porte sur les disciplines enseignées à l'Université, sur les capacités critiques et d'auto-évaluation des candidats, et elle requiert la référence aux exigences éthiques et aux règles déontologiques des psychologues.

Code signé par l'Association des Enseignants de Psychologie des Universités (AEPU), l'Association Nationale des Organisations de Psychologues (ANOP), la Société Française de Psychologie (SFP) le 22 mars 1996.

2. Recommandations internationales sur l'utilisation des tests [extrait][1]

Introduction et contexte d'origine

➤ Le besoin de Recommandations internationales

L'objectif essentiel visé par la Commission Internationale des Tests (en abrégé CIT) à travers ce projet de Recommandations est de promouvoir une bonne utilisation des tests et d'encourager des pratiques exemplaires dans le domaine de l'évaluation. Le travail réalisé jusqu'à maintenant par la CIT pour permettre un haut niveau de qualité dans l'adaptation des tests (Hambleton, 1994 ; Van de Vijver F. & Hambleton R., 1996) constitue une étape importante vers une homogénéisation de leur qualité, en vue de leur utilisation dans différentes langues et différentes cultures. Lors de sa réunion à Athènes en 1995, le conseil de la CIT a adopté une proposition visant à élargir cette préoccupation, en incluant des Recommandations sur une utilisation des tests qui soit équitable et conforme à l'éthique. De ces Recommandations peuvent être tirées des normes pour la définition des compétences des utilisateurs de tests et leur formation.

Il existe un certain nombre de raisons pour lesquelles le besoin de recommandations sur l'utilisation des tests au niveau international se manifeste :

- Les différents pays présentent des disparités importantes en ce qui concerne le niveau de contrôle légal, éventuel, qu'ils peuvent exercer sur l'utilisation du *testing* et ses conséquences pour ceux qui sont testés. Certaines organisations professionnelles nationales pratiquent un enregistrement légal des psychologues, d'autres non ; certaines disposent de procédures pour contrôler les normes d'utilisation des tests par des non-psychologues, d'autres n'en ont pas. L'existence d'un ensemble de recommandations, acceptées au niveau international, peut fournir aux associations nationales de psychologues et autres corps de professionnels concernés, une aide à la mise en place de normes, dans les pays où de telles normes sont soit actuellement déficientes, d'une manière ou d'une autre, soit totalement inexistantes.

1. Publié avec l'aimable autorisation de la Société Française de Psychologie (SFP). Le texte dans son entier est consultable sur le site de la SFP : www.sfpsy.org.

- L'organisation de l'accès aux tests, en termes de droits d'acquisition ou d'utilisation de ces instruments, varie de manière importante d'un pays à l'autre. Dans certains pays, cet accès est restreint aux seuls psychologues ; dans d'autres, aux utilisateurs répertoriés par les diffuseurs nationaux formellement autorisés ; dans d'autres encore, les utilisateurs peuvent accéder librement au matériel sans restriction auprès des diffuseurs dans leur pays ou directement auprès de diffuseurs domiciliés à l'étranger.
- Un certain nombre d'instruments bien connus sont apparus sur internet en violation des lois sur la propriété intellectuelle (*copyright*), sans l'autorisation des auteurs ou des éditeurs des tests, et sans considération pour les questions de sécurité des tests.
- Dans le domaine du *testing* en psychologie du travail, la plus grande mobilité internationale du travail a accru la demande de tests utilisables avec des candidats à un emploi venant de pays différents, les tests étant souvent administrés dans un pays pour le compte d'un employeur d'un autre pays.
- Un travail de développement est actuellement réalisé aux États-Unis et en Grande-Bretagne en vue de permettre une utilisation d'Internet pour une évaluation à distance dans les domaines professionnel et éducatif. Ce phénomène soulève une foule de questions concernant les normes d'administration et le contrôle du processus de *testing*, y compris le problème de la sécurité du test.

➤ But et objectifs

Le but à long terme de ce projet comprend la production d'un ensemble de recommandations qui se rapportent aux compétences (connaissances, capacités, savoir-faire et autres caractéristiques personnelles) requises des utilisateurs de tests. Ces compétences sont définies en termes de critères de performance évaluables. Ces critères fournissent la base pour développer des normes de compétence exigible de tout candidat à une qualification en tant qu'utilisateur de tests. L'analyse de telles compétences doit inclure la prise en compte de questions telles que :

- Les normes professionnelles et éthiques dans le *testing,*
- Les droits de la personne testée et des autres parties concernées par le processus de *testing*,
- Le choix et l'évaluation du test parmi un ensemble d'épreuves similaires,
- L'administration, la cotation et l'interprétation du test,

- Le compte rendu écrit et la communication des résultats.

Dans la mesure où elles sont directement liées à l'utilisation des tests, les Recommandations ont également des implications pour :

- Les normes à respecter pour la construction des tests,
- Les normes pour la documentation à l'usage des utilisateurs – par exemple, manuel de l'utilisateur, manuel technique,
- Les normes pour réguler l'achat et la disponibilité des tests, ainsi que l'information sur les tests.

Ces Recommandations représentent le travail de spécialistes dans le domaine du *testing* psychologique et éducatif (c'est-à-dire psychologues, psychométriciens, éditeurs de tests, auteurs de tests) issus d'un certain nombre de pays. L'intention de ce document n'est pas d'inventer de nouvelles recommandations, mais de rassembler les tendances communes qui parcourent les recommandations existantes, les codes de déontologie, les énoncés de normes et autres documents pertinents, pour créer une structure cohérente à l'intérieur de laquelle ces éléments puissent être compris et mis en œuvre.

➤ Mise au point des recommandations

Les Recommandations doivent être considérées comme des références par rapport auxquelles les normes locales existantes peuvent être comparées en ce qui concerne l'étendue de leur prise en charge, ainsi que leur qualité au niveau international. En utilisant les Recommandations comme références ou comme bases pour développer des documents valables localement (par exemple, normes, codes de déontologie, déclarations sur les droits des personnes testées), on favorisera l'accès à un haut niveau d'homogénéité transnationale.

Le travail sur les recommandations a débuté en rassemblant les documents se rapportant aux normes sur les tests, codes de déontologie, d'utilisation des tests, etc., dans un grand nombre de pays[1]. Bien que tirant parti de toutes ces sources, ces Recommandations ont été plus particulièrement influencées par :

1. Une liste de tous les documents qui ont alimenté ce processus peut être obtenue sur demande adressée aux auteurs.

- *The Australian Psychological Society (APS) Supplement to guidelines on the use of Psychological Tests* (Kendall et al., 1997).
- *The British Psychological Society (BPS) Level A and Level B standards for occupational test use* (Bartram, 1995, 1996).
- *The American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1985) Standards for educational and psychological testing.*
- *American Association for Counselling and Development (AACD) Responsibilities of Users of Standardized Tests (Schafer, W. D, 1992).*
- *The CPA (Canadian Psychological Association, 1987) Guidelines for Educational and Psychological Testing.*

Le document de l'*APS* a été précieux car il rassemble la plus grande partie de ce qui est contenu dans les publications de la *BPS* et les publications américaines, en tirant parti également des publications du *South African National Institute for Psychological Research (NIPR)*, et des conseils à l'intention des utilisateurs de tests publiés par les éditeurs de tests. Il intègre également beaucoup de ce qui provient des travaux fondateurs du *Joint Committee on Testing Practices (JCTP) Test User Qualifications Working Group's (TUQWG)*, travaux à partir d'une approche basée sur des données d'enquête pour promouvoir une bonne utilisation des tests (par exemple, Eyde et al., 1988, 1993 ; Morelandetal, 1995), et le travail du *JCTP* sur le *Code of Fair Testing Practices in Education* (JCTP, 1988 ; Fremer, Diamond, & Camara, 1989). L'annexe B a tiré davantage parti des travaux plus récents du *JCTP* (*JCTP*, 2000) sur les droits et responsabilités des personnes testées.

Le contenu des sources primaires a été analysé et les déclarations classées selon quatorze sections principales. Lorsque c'était approprié, des déclarations uniques ont été rédigées pour prendre en compte, en les synthétisant, un certain nombre de déclarations provenant de différentes sources. Les déclarations ont également été modifiées selon un format tel qu'il se présente comme la complémentation d'une phrase commune (par exemple, « Les utilisateurs de tests compétents feront tout leur possible pour... », ou « Les utilisateurs de tests compétents peuvent... »).

Cette structure initiale de quatorze sections principales et de leur contenu a été intégrée dans l'avant-projet de document de travail.

Celui-ci a constitué le matériel pour un atelier international qui s'est tenu à Dublin en juillet 1997. L'intention de l'atelier de la CIT était d'étudier et d'évaluer de manière critique tous les aspects du document de cadrage, en ayant pour objectif de produire une première version d'un ensemble

de recommandations qui auraient une crédibilité et une reconnaissance internationales. Pendant l'atelier, le document de cadrage a été examiné en détail, et des améliorations ont été proposées en termes de forme, de structure et de contenu. À la suite de l'atelier, le document a été complètement revu (version 2.0) et a circulé parmi tous ceux qui l'avaient commenté. Un avant-projet de document de consultation (version 3.1) a été préparé, qui prenait en compte tous les commentaires et suggestions proposés pour la version 2.0.

Des copies de la version 3.1 du document de consultation et une grille de réponse structurée ont été largement diffusées aux personnes et organisations-clé, pour commentaire. Un total de deux cents exemplaires a été distribué. Un total de vingt-huit réponses détaillées a été reçu, incluant des réponses d'organisations telles que l'APA, la BPS, la SFP et quelques autres associations professionnelles européennes. Durant l'été 1998, les Recommandations ont été revues à la lumière de ces commentaires, et 200 exemplaires (version 4.1) ont été envoyés pour une autre consultation. Un total de 18 réponses circonstanciées ont été reçues pour cette seconde série de consultations. De plus, des commentaires informels de soutien ont été fournis par de nombreux destinataires du document de consultation par courrier électronique ou lors de rencontres.

En mettant au point la présente version des Recommandations (Version 2000), tous les efforts ont été faits pour prendre en compte toutes ces réponses. Les réponses étaient, sans exception, utiles et constructives[1].

Ces Recommandations doivent être considérées comme une aide plutôt que comme une contrainte. Il est nécessaire de s'assurer que les Recommandations rassemblent les principes de base universels d'une pratique correcte des tests, sans chercher à imposer une uniformité là où existent des différences légitimes, d'un pays – ou d'une zone d'application – à l'autre, en ce qui concerne les fonctions ou les pratiques.

La structure proposée distingue trois principaux aspects des compétences :

1. Les normes professionnelles et éthiques de bonne pratique, qui concernent la façon selon laquelle le processus de *testing* est conduit, et la façon dont les utilisateurs de tests interagissent avec les autres personnes impliquées dans le processus.

1. Un compte rendu détaillé sur ces résultats de la première consultation a été soumis à la réunion du conseil de la CIT en août 1998. Un compte rendu de la seconde consultation joint à la Version 5.0 des recommandations a été soumis au conseil de la CIT lors de sa réunion de juin 1999. La Version 2000 contient des modifications de rédaction mineures par rapport à la Version 5.0.

2. Les connaissances, la compréhension et les savoir-faire relatifs au processus de *testing :* ce que les utilisateurs de test doivent être capables de faire.
3. Les connaissances et la compréhension qui sont nécessaires pour maîtriser le processus de *testing* – et l'étayer.

Ces trois composantes diffèrent, et sont pourtant inextricablement liées dans la pratique.

Les Recommandations proviennent d'un objectif-clé. Celui-ci peut être caractérisé comme « l'ordre de mission » de l'utilisateur de tests. Il constitue la base à partir de laquelle les Recommandations sont développées. Chaque recommandation définit l'une des facettes de compétence des utilisateurs de tests qui contribue à l'objectif clé.

Joint à l'objectif clé, le champ d'application des Recommandations décrit les personnes auxquelles elles s'appliquent, les formes d'évaluation auxquelles elles se rapportent et les contextes d'évaluation.

Ce document contient :

1. L'objectif clé et le champ d'application des Recommandations.
2. La définition des compétences des utilisateurs de tests, en relation avec une approche éthique des tests.
3. La définition des compétences des utilisateurs de tests, en relation avec une pratique correcte de l'utilisation des tests.

Les Recommandations

➤ Objectif-clé

Un utilisateur de tests compétent utilise les tests de manière appropriée, de manière professionnelle, et de manière éthique, en prenant en considération les besoins et les droits de ceux qui sont impliqués dans le processus de passation des tests, les justifications de la passation, et le contexte, au sens large, dans lequel la passation se déroule.

On permettra qu'il en soit ainsi en s'assurant que les utilisateurs de tests disposent des compétences nécessaires pour mener à bien une telle procédure, ainsi que les connaissances et une compréhension des tests et de leur utilisation suffisantes pour éclairer et étayer ce processus.

➤ Champ d'application

Toute tentative pour fournir une définition précise d'un test ou du *testing* en tant que processus échouera vraisemblablement parce qu'elle risque d'exclure certaines procédures qui devraient en faire partie, et d'en inclure d'autres qui devraient en être exclues. Pour les besoins de ces Recommandations, les termes tests et *testing* doivent être interprétés au sens large. Le fait qu'une procédure d'évaluation soit ou non qualifiée de test reste peu probant. Ces Recommandations sont pertinentes pour de nombreuses procédures d'évaluation qui ne sont pas appelées des tests ou pour lesquelles on cherche à éviter cette appellation. Plutôt que de fournir une définition unique, les propositions suivantes sont une tentative pour organiser le domaine couvert par les Recommandations.

- La passation de tests comprend une large gamme de procédures destinées à être employées dans l'évaluation psychologique, professionnelle et éducative.
- La passation de tests comprend des procédures permettant la mesure des comportements normaux ou pathologiques, voire des dysfonctionnements.
- Les procédures de passation de tests sont habituellement construites pour être administrées selon des conditions soigneusement contrôlées ou standardisées, qui incluent des protocoles cotés de manière systématique.
- Ces procédures fournissent des mesures de la performance et amènent à tirer des inférences à partir d'échantillons du comportement.
- Elles comprennent également des procédures qui peuvent aboutir à catégoriser ou à classer les personnes (par exemple, en termes de types psychologiques).

Toute procédure utilisée pour « tester », au sens défini ci-dessus, devrait être considérée comme un test, sans tenir compte de son mode d'administration, ni du fait qu'il a été, ou non, construit par un auteur de test professionnel, ni encore du fait qu'il comprendrait des ensembles de questions ou qu'il demande de réaliser des performances de tâches ou d'opérations. (par exemple, échantillon de tâches professionnelles, tests psychomoteurs de « poursuite »).

Les tests devraient s'appuyer sur des constats de leur fidélité et de leur validité en relation avec les objectifs poursuivis. Des preuves devraient être fournies pour appuyer les inférences qui sont tirées des scores au test. Ces preuves devraient être accessibles aux utilisateurs de tests, et disponibles

pour être examinées et évaluées de manière indépendante. Lorsque de telles informations importantes sont publiées dans des rapports techniques difficilement accessibles, des résumés comprenant les références complètes devraient être fournis par le diffuseur du test.

Les Recommandations sur l'utilisation des tests doivent être considérées comme s'appliquant à toutes les procédures semblables qu'elles soient ou non désignées explicitement comme « tests psychologiques » ou « tests éducatifs » et qu'elles soient ou non confirmées par des constats techniques disponibles.

La plupart de ces Recommandations s'appliqueront aussi à des procédures d'évaluation situées en dehors du domaine des tests. Elles peuvent être pertinentes pour toute procédure d'évaluation utilisée dans des situations où l'évaluation des personnes se fait dans un but sérieux et significatif et qui, mal utilisée, pourrait aboutir à des dommages aux personnes ou à des souffrances psychologiques (par exemple, les entretiens de sélection professionnelle, les évaluations des performances professionnelles, l'évaluation diagnostique des besoins d'aide aux apprentissages scolaires/cognitifs).

Les Recommandations ne s'appliquent pas à l'utilisation de matériels qui peuvent avoir une ressemblance superficielle avec les tests, mais que tous les participants reconnaissent comme destinés à être utilisés seulement à des fins de distraction ou d'amusement (par exemple, questionnaires de style de vie dans les magazines et les journaux).

➤ À qui s'adressent les Recommandations ?

Les Recommandations s'appliquent à l'utilisation des tests dans une pratique professionnelle. En tant que telles, elles s'adressent d'abord :

- aux personnes qui acquièrent ou qui détiennent des matériels de test ;
- à ceux qui ont la responsabilité de choisir des tests et de déterminer l'usage qui en sera fait ;
- à ceux qui administrent, cotent ou interprètent les tests ;
- à ceux qui fournissent des avis aux autres sur la base des résultats aux tests (par exemple, psychologues cliniciens, psychologues du travail, psychologues scolaires, conseillers d'orientation-psychologues, etc.) ;
- à ceux qui ont la charge de rendre compte des résultats aux tests et de communiquer leurs résultats aux personnes qui ont passé des tests.

Les Recommandations sont également pertinentes pour d'autres personnes impliquées dans l'utilisation des tests telle qu'elle a été définie ci-dessus. Celles-ci comprennent :

- les constructeurs (auteurs) de tests,
- les éditeurs de tests,
- ceux qui sont partie prenante dans la formation des utilisateurs de tests,
- ceux qui sont testés, ainsi que leur entourage (parents, épouse, partenaires de vie),
- les organisations professionnelles et les autres associations qui sont concernées par l'utilisation des tests psychologiques et éducatifs,
- les décideurs et les législateurs.

Bien que destinées au départ aux pratiques professionnelles, les Recommandations seront également pertinentes pour ceux qui utilisent les tests uniquement à des fins de recherche.

Les Recommandations n'ont pas pour but de couvrir tous les types de techniques d'évaluation (par exemple, les entretiens structurés ou semi-structurés, l'évaluation des activités de groupe), ou toutes les situations dans lesquelles une évaluation a lieu (par exemple, les centres d'évaluation pour l'emploi [*assessment centers*]). Cependant, plusieurs des Recommandations peuvent vraisemblablement s'appliquer dans des situations d'évaluation et pour des objectifs plus généraux que ceux observés en premier lieu dans le *testing* psychologique et éducatif (par exemple, l'utilisation des centres de bilan pour le placement ou la sélection des salariés, les entretiens structurés ou semi-structurés, ou l'évaluation pour la sélection, l'orientation professionnelle et le conseil en carrière).

➤ Facteurs contextuels

Les Recommandations s'appliquent au niveau international. Elles peuvent être utilisées pour développer des normes spécifiques et locales (par exemple, nationales) en passant par un processus de contextualisation. Il est admis que de nombreux facteurs affectent la manière dont les normes de qualité peuvent être gérées et mises en place dans la pratique. Ces facteurs contextuels doivent être pris en considération au niveau local (national) lorsqu'on interprète les Recommandations et qu'on cherche à définir ce qu'elles veulent dire de manière pratique dans un environnement particulier.

Les facteurs qui doivent être pris en considération, lorsqu'on transforme les Recommandations en normes spécifiques, comprennent :

- les différences sociales, politiques, institutionnelles, linguistiques et culturelles entre les cadres d'évaluation ;
- les lois des pays où se déroule le *testing ;*
- les Recommandations nationales existantes et les normes de qualité élaborées par des associations et des organisations professionnelles de psychologues ;
- les différences se rapportant aux évaluations individuelles et aux évaluations de groupe ;
- les différences se rapportant au domaine du test (éducatif, clinique, travail et autres champs d'évaluation) ;
- les principaux destinataires des résultats des tests (par exemple les personnes testées, leurs parents ou leur tuteur, le commanditaire du test, un employeur ou un tiers) ;
- les différences relatives à l'utilisation des résultats du test (e.g., pour prendre une décision à l'issue d'un examen de sélection, ou pour fournir des informations dans le cadre d'une activité de conseil) ;
- les variations dans le degré auquel la situation fournira la possibilité de vérifier l'exactitude de l'interprétation à la lumière d'informations ultérieures et de la modifier si nécessaire.

➤ Connaissances, compréhension et savoir-faire

Connaissances, compréhension et savoir-faire étayent toutes les compétences des utilisateurs de tests. La nature de leur contenu et leur niveau de détail peuvent varier selon les pays, les domaines d'application, et le niveau de compétence requis pour utiliser un test.

Les Recommandations ne comportent pas de description détaillée de ces éléments. Cependant, lorsqu'on applique les Recommandations dans des situations spécifiques, les connaissances pertinentes, les aptitudes, compétences et autres caractéristiques personnelles devront être spécifiées. Cette spécification fait partie du processus de contextualisation par lequel des recommandations génériques sont développées dans des normes spécifiques. Les descriptions des principaux domaines de connaissances, compréhension, savoir-faire doivent comprendre les points suivants.

Connaissances déclaratives pertinentes

- Connaissance des principes et procédures de base de la psychométrie, et des exigences techniques des tests (par exemple, fidélité, validité, standardisation) ;
- Connaissance suffisante des tests et de la mesure, pour permettre une compréhension appropriée des résultats des tests ;
- Connaissance et compréhension des théories pertinentes et des modèles des aptitudes, de la personnalité et d'autres construits psychologiques ou de la psychopathologie, autant que nécessaire pour s'informer sur le choix des tests et l'interprétation des résultats ;
- Connaissance des tests et des fournisseurs de tests dans le secteur d'application où on intervient.

Connaissances pratiques et compétences

- Connaissances et compétences relatives aux procédures spécifiques d'évaluation ou aux instruments, y compris l'utilisation des procédures d'évaluation assistée par ordinateur ;
- Connaissances spécialisées et compétences pratiques nécessaires pour une bonne utilisation des tests situés à l'intérieur du répertoire d'outils d'évaluation de chacun ;
- Connaissances et compréhension de la – ou des – théorie(s) sous-jacente(s) aux scores au test, lorsque c'est important si l'on veut être en mesure de tirer des inférences valides à partir des résultats au test.

➤ **Les recommandations couvrent :**

Des compétences générales et personnelles relatives aux tâches

- La réalisation d'activités pertinentes telles que l'administration des tests, le compte rendu et la préparation de la communication des résultats aux personnes testées et aux autres clients ;
- Des compétences suffisantes en communication écrite et orale pour une préparation appropriée des personnes testées, l'administration des tests, la rédaction de comptes rendus des résultats aux tests, et pour interagir avec les autres personnes concernées (parents, ou décideurs dans les organisations) ;
- Des compétences relationnelles suffisantes pour une préparation appropriée des personnes testées, l'administration des tests, et la préparation de la communication des résultats.

Des connaissances et compétences contextuelles

- Savoir quand utiliser ou ne pas utiliser les tests ;
- Savoir comment intégrer le *testing* avec d'autres composantes moins formelles de la situation d'évaluation (par exemple données biographiques, entretiens non structurés et références, etc.) ;
- Connaissance des questions d'actualité professionnelle, légale et éthique concernant l'utilisation des tests, et de leurs implications pratiques pour l'utilisation des tests.

Des savoir-faire dans la gestion des tâches

- Connaissance des règles de déontologie et de pratique correcte concernant l'utilisation des tests et de leurs résultats, la préparation d'un compte rendu, sa production et son archivage, le stockage en sécurité des matériels de tests et des données de tests ;
- Connaissance des contextes sociaux, culturels et politiques dans lesquels le test est utilisé, et des modalités selon lesquelles ces facteurs peuvent avoir un effet sur les résultats, leur interprétation et l'utilisation qui en est faite.

Des compétences quant à la gestion des imprévus

- Savoir comment gérer les problèmes, difficultés et interruptions en cours de déroulement ;
- Savoir comment gérer les questions posées par une personne testée pendant l'administration du test, etc.
- Savoir comment gérer des situations dans lesquelles il existe une possibilité de mauvais usage des tests ou un risque de mauvaise interprétation des scores au test.

Prendre ses responsabilités pour un usage éthique des tests

➤ **Les utilisateurs de tests compétents devraient :**

1.1. Agir de façon professionnelle et éthique.

1.1.1. Promouvoir et maintenir des normes professionnelles et éthiques.

1.1.2. Être capables de mettre en pratique une compréhension des questions professionnelles et éthiques actuelles et des débats concernant l'utilisation des tests et leur champ d'application.

1.1.3. Mettre en place un système de règles explicite sur le *testing* et l'utilisation des tests[1].

1.1.4. S'assurer que les personnes travaillant pour, ou avec eux, adhèrent aux normes éthiques et déontologiques.

1.1.5. Diriger les communications de résultats en prenant en compte les sensibilités des personnes testées et des tierces parties concernées.

1.1.6. Présenter les tests et le *testing* de façon positive et équitable dans les communications avec et à partir des médias.

1.1.7. Éviter les situations où ils peuvent avoir – ou être perçus comme ayant – un intérêt personnel dans les résultats de l'évaluation, ou dans lesquelles l'évaluation risque de nuire à la relation avec leur client.

1.2. S'assurer qu'ils ont les compétences pour utiliser les tests.

1.2.1. Travailler dans les limites des principes scientifiques et de l'expérience établie.

1.2.2. Atteindre et maintenir un haut niveau d'exigences quant à leurs compétences personnelles.

1.2.3. Connaître les limites de leurs propres compétences et travailler à l'intérieur de ces limites.

1.2.4. Suivre les évolutions pertinentes et les progrès concernant l'utilisation des tests, et le développement des tests, y compris les changements législatifs et politiques qui peuvent avoir un impact sur les tests et l'utilisation des tests.

1.3. Prendre leurs responsabilités pour l'utilisation qu'ils font des tests.

1.3.1. Ne proposer que les activités de *testing* et n'utiliser que les tests pour lesquels ils sont qualifiés.

1.3.2. Assumer ses responsabilités pour le choix des tests utilisés et pour les conseils formulés.

1.3.3. Fournir, aux participants au processus de *testing,* des informations claires et adéquates sur les règles d'éthique et les dispositions légales régissant le *testing* psychologique.

1.3.4. S'assurer que le contenu du contrat entre la personne testée et la personne qui fait passer les tests est clair et bien compris[2].

1.3.5. Être vigilant pour détecter toute conséquence inattendue de l'usage des tests.

1. Un exemple d'ébauche de système de règles est présenté en annexe A.

2. On trouvera un exemple de « contrat » entre la personne testée et la personne faisant passer les tests en annexe B.

1.3.6. Faire tout son possible pour éviter de nuire ou de causer une souffrance à ceux qui sont impliqués dans le processus de test.

1.4. S'assurer que le matériel de test est conservé en sécurité.

1.4.1. Sécuriser le stockage du matériel de test et en contrôler l'accès.

1.4.2. Respecter les lois sur la propriété intellectuelle et les accords qui existent en ce qui concerne le test, incluant les interdictions de reproduction, ou la transmission du matériel au format électronique – ou autre – à d'autres personnes, que celles-ci soient ou non qualifiées.

1.4.3. Protéger l'intégrité des tests en s'abstenant de donner un entraînement aux sujets sur du matériel de test ayant cours, ou un autre matériel d'entraînement dont l'usage pourrait influencer de manière inéquitable leurs performances aux tests.

1.4.4. S'assurer que les techniques de tests ne sont pas décrites publiquement d'une façon telle que leur utilité en soit affectée.

1.5. S'assurer que les résultats aux tests sont traités confidentiellement.

1.5.1. Préciser qui aura accès aux résultats et définir des niveaux de confidentialité.

1.5.2. Expliquer les niveaux de confidentialité aux personnes avant que les tests ne soient administrés.

1.5.3. Limiter l'accès aux résultats à ceux qui y sont autorisés.

1.5.4. Obtenir un consentement éclairé avant de communiquer les résultats à d'autres personnes.

1.5.5. Protéger les données stockées sur fichier électronique de telle manière que seules les personnes autorisées puissent y accéder.

1.5.6. Établir des règles claires concernant la durée pendant laquelle les données de tests sont conservées dans des fichiers.

1.5.7. Ôter les noms et autres identifiants personnels des bases de données contenant des résultats qui sont archivés à des fins de recherches, d'élaboration de normes (étalonnages), ou d'autres traitements statistiques.

Assurer une pratique correcte dans l'utilisation des tests

2.1. Estimer l'intérêt éventuel d'une utilisation des tests dans une situation d'évaluation donnée.

Les utilisateurs de tests compétents devront :

2.1.1. Produire une justification argumentée de l'utilisation de tests.

2.1.2. S'assurer qu'il a été procédé à une analyse approfondie des besoins du client, des motifs de la consultation, ou du type de diagnostic, de situation, ou d'emploi visé par cette évaluation.

2.1.3. Établir que les connaissances, savoir-faire, compétences, aptitudes ou autres caractéristiques, que le test est censé mesurer, sont des indicateurs des comportements pertinents dans le contexte à partir duquel on fera des inférences.

2.1.4. Rechercher d'autres sources collatérales d'informations pertinentes.

2.1.5. Estimer les avantages et les inconvénients de l'utilisation de tests, par comparaison avec d'autres sources d'informations.

2.1.6. S'assurer qu'un plein usage est fait de toutes les sources d'informations collatérales.

2.2. Choisir des tests techniquement fiables et appropriés à la situation.

Les utilisateurs de tests compétents devront :

2.2.1. Examiner l'information actualisée couvrant l'ensemble des tests potentiellement pertinents (par exemple à partir de jeux de spécimens, d'études indépendantes, de conseils d'experts), avant de choisir un test à utiliser.

2.2.2. Déterminer si la documentation technique et le manuel de l'utilisateur fournissent des informations suffisantes pour apprécier les points suivants :

a) Portée ou couverture et représentativité du contenu du test, pertinence des étalonnages, niveau de difficulté du contenu, etc. ;

b) Précision de la mesure et fidélité démontrées en ce qui concerne les populations de références ;

c) Validité (en ce qui concerne les populations de référence) et pertinence pour l'usage requis ;

d) Absence de biais systématiques au détriment de l'un des groupes de sujets auxquels le test sera administré ;

e) Caractère acceptable pour ceux qui seront impliqués dans son utilisation, prenant en compte l'équité et la pertinence perçues ;

f) Faisabilité, tenant compte de la durée, du coût et des besoins en général.

2.2.3. Se garder de l'utilisation de tests qui ont une documentation technique inadaptée ou peu claire.

2.2.4. N'utiliser des tests que dans les situations pour lesquelles des preuves de validité pertinentes et appropriées sont disponibles.

2.2.5. Se garder de porter un jugement sur un test seulement sur la base de sa validité apparente, des témoignages des utilisateurs, ou du conseil de personnes qui y ont des intérêts commerciaux.

2.2.6. Répondre aux demandes de toutes les parties (par exemple, les personnes testées, les parents, les responsables hiérarchiques), en leur fournissant une information suffisante pour leur permettre de comprendre pourquoi le test a été choisi.

2.3. Prendre effectivement en compte les questions d'équité dans l'utilisation des tests.

Lorsqu'on utilise des tests avec des personnes appartenant à différents groupes (par exemple, en termes de sexe, d'origine culturelle, d'éducation, ou d'âge), les utilisateurs de tests compétents s'assureront, autant que possible, que :

2.3.1. Les tests ne sont pas biaisés et sont adaptés pour les différents groupes qui vont être testés.

2.3.2. Les dimensions qui sont évaluées, sont significatives dans chacun des groupes en présence.

2.3.3. Des données sont disponibles sur l'existence de différences possibles dans les performances des groupes au test.

2.3.4. Des constats concernant le Fonctionnement Différentiel des Items (FDI[1]) sont disponibles, lorsque c'est pertinent.

2.3.5. On dispose de données confirmant la validité du test, compte tenu de son utilisation prévue pour les différents groupes.

2.3.6. Les effets des différences intergroupes non pertinentes par rapport à l'objectif principal de l'évaluation (par exemple différences de motivation pour répondre, ou compétences en lecture) sont minimisés.

2.3.7. Dans tous les cas, les Recommandations concernant l'usage équitable des tests sont interprétées à la lumière du contexte des politiques et des législations nationales.

Lorsque les tests utilisés sont administrés dans plusieurs langues (à l'intérieur d'un même pays ou entre plusieurs pays), les utilisateurs de tests compétents s'assureront, autant que possible, que :

1. Note des traducteurs : Le FDI est traditionnellement appelé biais d'item ou biais item/test. le FDI se manifeste lorsqu'un item mesure une autre variable que la variable qu'il est censé mesurer et que cette variable « parasite » favorise – ou défavorise – un des groupes en présence. Une nuisance est ainsi introduite dans la mesure. Pour une revue de questions récente sur les biais dans les tests et le FDI, on peut consulter Vrignaud, P. (2002). Les biais de mesure : savoir les identifier pour y remédier. *Bulletin de Psychologie*, *55(6)*, 625-634.

2.3.8. La version dans chacune des langues ou dialectes a été mise au point selon une méthodologie rigoureuse et répondant à un niveau d'exigence de qualité élevé.

2.3.9. Les constructeurs ont été attentifs aux questions de contenu, de culture et de langue.

2.3.10. Ceux qui administreront les tests peuvent communiquer clairement dans la langue dans laquelle le test doit être administré.

2.3.11. Le niveau de compétence des sujets, pour la langue dans laquelle le test sera administré, est contrôlé de manière systématique, et, selon ce qui est le plus adéquat, le sujet est évalué avec une version du test dans sa langue ou selon une procédure bilingue.

Quand on prévoit d'utiliser les tests avec des personnes handicapées, les utilisateurs de tests compétents s'assureront, autant que possible, que :

2.3.12. On a recherché les avis d'experts compétents concernant les effets potentiels des différents handicaps sur la performance aux tests.

2.3.13. On a demandé leur avis aux personnes susceptibles de passer le test, et leurs besoins et souhaits sont pris en considération de manière appropriée.

2.3.14. Les aménagements adéquats ont été prévus lorsqu'il y a parmi les personnes testées des personnes ayant des difficultés d'audition, de vision, de motricité, ou d'autres handicaps (par exemple, difficultés d'apprentissage, dyslexie).

2.3.15. L'utilisation d'autres instruments d'évaluation, plutôt que des modifications des tests eux-mêmes, a été envisagée (par exemple, d'autres tests plus adaptés, ou d'autres formes structurées d'évaluation).

2.3.16. L'avis de spécialistes compétents a été demandé si l'importance des modifications requises pour l'utilisation avec les personnes handicapées dépasse l'expérience de l'utilisateur de tests.

2.3.17. Les modifications, si nécessaires, sont adaptées à la nature du handicap et sont mises en œuvre pour minimiser son impact sur la validité des scores.

2.3.18. Les informations concernant la nature de toutes les modifications faites à un test ou à une procédure de test sont communiquées à ceux qui interprètent ou travaillent à partir des scores aux tests, chaque fois que la rétention d'une telle information pourrait conduire à une interprétation biaisée ou à une décision inéquitable.

2.4. Faire les préparations requises pour la séance de tests.

Les utilisateurs de tests compétents devront faire tous les efforts raisonnables pour être sûrs de :

2.4.1. Fournir aux parties concernées, en temps opportun, une information claire concernant l'objectif de l'utilisation de tests, la façon dont ils peuvent le mieux se préparer à la séance de tests et la procédure à suivre.

2.4.2. Informer les personnes testées, de la langue ou du dialecte pour lesquels le test est considéré comme approprié.

2.4.3. Envoyer aux personnes testées des exercices d'entraînement, échantillons, ou documents de préparation, lorsque ceux-ci sont disponibles et lorsqu'une telle pratique est cohérente avec les usages recommandés pour les tests concernés.

2.4.4. Expliquer clairement aux personnes testées leurs droits et leurs responsabilités[1].

2.4.5. Recueillir l'accord explicite des personnes testées ou de leurs représentants légaux avant toute administration de test.

2.4.6. Expliquer aux parties concernées, lorsque la passation des tests est facultative, les conséquences d'une acceptation ou d'un refus de passer les tests, de sorte qu'elles puissent faire leur choix en connaissance de cause.

2.4.7. Effectuer les aménagements matériels nécessaires en s'assurant que :

a) Les préparatifs sont conformes à ceux qui sont prescrits dans le manuel de l'éditeur.

b) Les lieux et les installations pour la passation des tests ont été préparés suffisamment à l'avance, l'environnement physique est accessible, sûr, tranquille, ne gênant pas la concentration, et approprié à l'objectif visé.

c) Les documents, en nombre suffisant, sont disponibles et ont été vérifiés afin de s'assurer qu'aucune trace n'a été laissée par les utilisateurs précédents sur les livrets de questions ou sur les feuilles de réponse.

d) Le personnel qui sera impliqué dans l'administration est compétent ;

e) Des aménagements appropriés ont été prévus pour tester les personnes présentant un handicap.

2.4.8. Anticiper les problèmes possibles et y remédier par une préparation minutieuse du matériel et des instructions.

2.5. Administrer les tests de manière appropriée.

Les utilisateurs de tests compétents devraient :

2.5.1. Établir un climat favorable en accueillant les personnes à tester et en les informant de manière positive.

1. Voir annexe B.

2.5.2. Agir pour réduire l'anxiété des personnes testées et éviter de créer ou de renforcer une anxiété inutile.

2.5.3. S'assurer que les facteurs de distraction potentiels (par exemple, les alarmes de montre, les téléphones portables, les bippeurs) ont été neutralisés.

2.5.4. S'assurer avant le début de la séance que les personnes testées ont en leur possession le matériel nécessaire pour passer le test.

2.5.5. Administrer les tests selon des conditions de surveillance appropriées.

2.5.6. Dans la mesure du possible, donner les consignes du test dans la langue principale des personnes testées, même quand le contenu du test a été conçu pour fournir des informations sur les connaissances et les compétences dans une seconde langue.

2.5.7. Suivre strictement les indications et les instructions telles qu'elles sont spécifiées dans le manuel du test, et prévoir des aménagements raisonnables pour les personnes handicapées.

2.5.8. Lire les instructions clairement et calmement.

2.5.9. Laisser assez de temps pour terminer les exemples.

2.5.10. Observer et noter les divergences par rapport à la procédure de passation du test.

2.5.11. Surveiller et noter les temps de réponse avec précision, lorsque c'est prévu dans la procédure.

2.5.12. S'assurer que tout le matériel a été récupéré à la fin de chaque passation de tests.

2.5.13. Administrer les tests en s'assurant d'un niveau adéquat de surveillance et d'authentification de l'identité des personnes testées.

2.5.14. S'assurer que ceux qui aident à l'administration des tests ont reçu une formation appropriée.

2.5.15. S'assurer que les personnes testées ne restent pas sans surveillance ou que des facteurs extérieurs ne les distraient pendant une séance de tests surveillée.

2.5.16. Fournir une assistance appropriée aux personnes testées qui montrent des signes de détresse ou d'anxiété excessifs.

2.6. Corriger et analyser les tests avec exactitude. Les utilisateurs de tests compétents devront :

2.6.1. Se conformer strictement aux procédures standardisées pour établir les scores.

2.6.2. Effectuer la transformation appropriée des notes brutes en d'autres types d'échelles pertinentes.

2.6.3. Choisir des types d'échelles appropriés à l'usage que l'on se propose de faire des scores au test.

2.6.4. Vérifier l'exactitude de la conversion des scores en d'autres échelles et de toutes les autres procédures de calcul.

2.6.5. S'assurer que des conclusions invalides ne sont pas tirées de la comparaison de scores avec des normes inadaptées aux personnes testées, ou périmées.

2.6.6. Calculer, lorsque c'est approprié, des scores composites en utilisant les formules et les équations standards.

2.6.7. Mettre en œuvre des procédures pour repérer des scores improbables ou aberrants parmi les résultats des tests.

2.6.8. Porter clairement et précisément les noms des échelles dans les comptes rendus et fournir des éléments d'information clairs sur les normes, les types d'échelles et les équations utilisées.

2.7. Interpréter les résultats de manière appropriée.

Les utilisateurs de tests compétents devraient :

2.7.1. Maîtriser la compréhension des fondements théoriques et conceptuels du test, la documentation technique, et les directives pour l'utilisation et l'interprétation des échelles.

2.7.2. Bien comprendre les échelles utilisées, les caractéristiques des normes ou des groupes de référence et les limites des scores.

2.7.3. Prendre des mesures pour minimiser les effets sur l'interprétation du test des biais éventuels que l'utilisateur pourrait introduire à l'encontre des membres du groupe culturel auquel appartient la personne testée.

2.7.4. Utiliser des normes ou des groupes de référence appropriés lorsqu'ils sont disponibles.

2.7.5. Interpréter les résultats à la lumière des informations disponibles sur les personnes testées (par exemple, l'âge, le sexe, le niveau d'éducation, la culture et autres facteurs) en prenant en compte, de manière adéquate, les limitations techniques du test, du contexte d'évaluation, et des besoins de ceux qui ont un intérêt légitime dans les résultats du processus.

2.7.6. Éviter de généraliser à outrance les résultats d'un test jusqu'à des traits ou des caractéristiques humaines qui ne sont pas mesurées par le test.

2.7.7. Prendre en considération, lorsqu'on interprète les scores, la fidélité de chaque échelle, l'erreur de mesure et autres caractéristiques qui ont pu modifier artificiellement les scores.

2.7.8. Prendre en compte les critères de validité, concernant la variable mesurée, pour les membres du groupe démographique auquel appartient la personne testée (par exemple, groupe culturel, âge, classe sociale, et sexe).

2.7.9. Dans l'interprétation des tests, n'utiliser des scores d'admission que si des preuves de la validité de ces scores d'admission sont disponibles et valident leur utilisation.

2.7.10. Être attentif aux stéréotypes sociaux se rapportant au groupe auquel appartient la personne testée (par exemple, groupe culturel, âge, classe sociale, et sexe) et éviter d'interpréter le test d'une façon qui perpétue de tels stéréotypes.

2.7.11. Prendre en compte, au niveau du groupe ou de l'individu, toute déviation de la procédure standard dans l'administration du test.

2.7.12. Prendre en compte tout indice d'une familiarisation antérieure avec le test lorsqu'il existe des données disponibles concernant l'effet d'une telle familiarisation sur la performance au test.

2.8. Communiquer les résultats clairement et précisément aux personnes concernées.

Les utilisateurs de tests compétents devraient :

2.8.1. Identifier les parties à qui l'on peut, de manière légitime, communiquer les résultats aux tests.

2.8.2. Avec le consentement éclairé des personnes testées, ou de leurs représentants légaux, produire des comptes rendus écrits ou oraux pour les parties intéressées.

2.8.3. S'assurer que les niveaux de technicité et de langage sont adaptés au niveau de compréhension des destinataires.

2.8.4. Souligner le fait que les résultats des tests ne sont qu'une source d'information et doivent toujours être considérés en liaison avec d'autres types d'information.

2.8.5. Expliquer comment l'importance des résultats aux tests doit être pondérée en relation avec les autres informations sur la personne évaluée.

2.8.6. Utiliser pour le rapport une présentation et un plan qui soient appropriés au contexte de l'évaluation.

2.8.7. Quand cela s'avère opportun, fournir aux décideurs des informations sur la manière dont les résultats peuvent être utilisés pour éclairer leur décision.

2.8.8. Expliquer et argumenter l'utilisation des résultats aux tests ayant servi pour classer les personnes en catégories (par exemple, à des fins de diagnostic ou de sélection professionnelle).

2.8.9. Introduire dans les rapports écrits des résumés clairs, et, lorsque c'est pertinent, des recommandations spécifiques.

2.8.10. Donner un compte rendu oral aux personnes testées qui soit constructif et, puisse les aider.

2.9. Contrôler l'adéquation du test, et de son utilisation.

Les utilisateurs de tests compétents devraient :

2.9.1. Contrôler et passer périodiquement en revue les changements qui se sont produits au cours du temps dans les populations testées, et dans tous les critères utilisés.

2.9.2. Vérifier si les tests n'ont pas d'éventuels impacts négatifs.

2.9.3. Être attentifs à la nécessité de réévaluer l'utilisation d'un test si des changements sont apportés à sa forme, à son contenu ou à son mode d'administration.

2.9.4. Être attentifs à l'opportunité de réévaluer les preuves de validité du test si l'objectif pour lequel il est utilisé, a été modifié.

2.9.5. Lorsque c'est possible, chercher à valider les tests pour l'usage qui en est fait, ou participer à des études de validation systématiques.

2.9.6. Lorsque c'est possible, contribuer à la mise à jour des informations concernant les normes, la fidélité, et la validité du test, en transmettant aux constructeurs du test, éditeurs ou chercheurs, des données pertinentes.

Bibliographie

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing*. Washington DC : American Psychological Association.

Bartram, D. (1995).The Development of Standards for the Use of Psychological Tests in Occupational Settings : The Competence Approach. *The Psychologist*, May, 219-223.

Bartram, D. (1996). Test Qualifications and Test Use in the UK : The Competence Approach. *European Journal of Psychological Assessment*, 12, 62-71.

Canadian Psychological Association. (1987). *Guidelines for Educational and Psychological Testing.* Ottawa : Canadian Psychological Association.

Eyde, L.D., Moreland, K.L. & Robertson, G.J. (1988). *Test User Qualifications : A Data-based Approach to Promoting Good Test Use.* Report for the Test User Qualifications Working Group. Washington DC : American Psychological Association.

Eyde, L.D., Robertson, G.J., Krug, S.E. et al (1993). *Responsible Test Use : Case Studies For Assessing Human Behaviour.* Washington DC : American Psychological Association.

Fremer, J., Diamond, E.E. & Camara, W.J. (1989). Developing a Code of Fair *Testing* Practices in Education. *American Psychologist*, 44, 1062-1067.

Hambleton, R. (1994).Guidelines for adapting educational and psychological tests : A progress report. *European Journal of Psychological Assessment*, 10, 229-244.

Joint Committee on *Testing* Practices. (1988). *Code of Fair Testing Practices in Education*. Washington DC : Joint Committee on *Testing* Practices.

Joint Committee on Testing Practices. (2000). Rights and Responsibilities of Test Takers : Guidelines and Expectations. Washington DC : Joint Committee on Testing Practices.

Kendall, I., Jenkinson, J., De Lemos, M. & Clancy, D. (1997). *Supplement to Guidelines for the use of Psychological Tests.* Australian Psychological Society.

Moreland, K.L., Eyde, L.D., Robertson, G.J., Primoff, E.S. & Most, R.B. (1995). Assessment of Test User Qualifications : A Research-Based Measurement Procedure. *American Psychologist*, 50, 14-23.

Schafer, W.D. (1992). *Responsibilities of Users of Standardized Tests : RUST Statement Revised.* Alexandria, VA : American Association for Counseling and Development.

Van de Vijver, F. & Hambleton, R. (1996). Translating tests : some practical guidelines. *European Psychologist*, 1, 89-99.

Annexes

➤ Annexe A : Recommandations pour l'ébauche d'une politique sur le testing

Les recommandations suivantes concernent le besoin que peuvent avoir les organisations de réfléchir, de manière systématique, à leur politique de *testing* et de s'assurer que toute personne concernée a une idée claire de cette politique. Le besoin d'une politique explicite de *testing* n'est pas limité aux grandes organisations. Les PME et les PMI qui utilisent les tests, aussi bien que les grandes sociétés, devraient être attentives à leur politique de *testing*, de la même manière qu'elles le sont aux questions de santé et de sécurité, à la parité, aux handicaps et autres domaines considérés dans le cadre des pratiques correctes de gestion et de traitement du personnel[1].

1. NDT : En France, la loi du 31 décembre 1992, encadre les pratiques d'évaluation en milieu professionnel.

Bien que les considérations et les obligations suivantes puissent être aménagées pour être mises en œuvre par les utilisateurs de tests travaillant comme praticiens indépendants, il n'en est pas moins important que ceux-ci aient une bonne compréhension de leur propre politique et qu'ils doivent savoir la communiquer à leurs partenaires.

Une politique sur le *testing* est élaborée afin de :

- S'assurer que les objectifs des personnes et des organisations sont atteints ;
- S'assurer que les mauvais usages potentiels sont évités ;
- Montrer son engagement envers les pratiques correctes ;
- S'assurer que l'usage des tests est approprié au but poursuivi ;
- S'assurer que les tests ne produisent pas de discriminations inéquitables ;
- S'assurer que les évaluations sont basées sur des informations complètes et pertinentes ;
- S'assurer que les tests ne sont utilisés que par des personnels qualifiés.

Une politique sur le *testing* devrait couvrir la plupart, sinon toutes, les questions suivantes :

- Utilisation appropriée des tests ;
- Mise en sécurité des tests et des protocoles ;
- Qui peut administrer, coter, et interpréter les tests ;
- Les conditions de qualification pour ceux qui veulent utiliser les tests ;
- La formation des utilisateurs de tests ;
- La préparation des personnes testées ;
- L'accès au matériel et sa mise en sécurité ;
- L'accès aux résultats des tests et aux éléments confidentiels des protocoles ;
- La communication des résultats aux personnes testées ;
- La responsabilité envers les personnes testées avant, pendant et après la passation des tests ;
- Les responsabilités personnelles et institutionnelles de chacune des personnes utilisatrices.

Toute politique doit être revue régulièrement et mise à jour lorsque des évolutions dans le *testing* ou des changements dans les pratiques ont lieu.

Les parties concernées doivent pouvoir avoir accès à la politique de *testing* et en être informés.

La responsabilité de la politique de *testing* de toute organisation devrait être déposée auprès d'un utilisateur de test qualifié disposant de l'autorité pour s'assurer de la mise en place et du respect de cette politique.

➤ Annexe B : Recommandations pour établir des relations contractuelles entre les parties concernées par le processus de testing.

Les contrats entre l'utilisateur de test et les personnes testées devraient être cohérents avec les pratiques correctes, la législation et la politique sur le *testing* de l'utilisateur de test. Les éléments suivants sont donnés comme exemple de la nature des questions qu'un tel contrat devrait aborder. Les détails peuvent varier en fonction du contexte de l'évaluation (par exemple, travail, éducation, clinique, recherche) et des lois et dispositions réglementaires locales ou nationales.

Les contrats entre les utilisateurs de tests, les personnes testées, et les autres parties, sont souvent implicites et tacites (au moins partiellement). Clarifier les attentes, les rôles et les responsabilités de toutes les parties peut permettre d'éviter les malentendus, les dommages et les litiges.

Pour sa part, l'utilisateur de test fera tout son possible pour :

b. 1. Informer les personnes testées de leurs droits en ce qui concerne la manière dont leurs résultats aux tests seront utilisés et de leurs droits d'y avoir accès[1].

b. 2. Donner un avertissement a priori et précis sur tous les coûts financiers que peut entraîner le processus de *testing*, qui sera responsable du paiement et la date à laquelle ce paiement sera exigible.

b. 3. Traiter les personnes testées avec courtoisie, respect et impartialité sans distinction d'origine ethnique, de sexe, d'âge, de handicap, etc.

b. 4. Utiliser des tests fiables, adaptés aux personnes testées et aux objectifs de l'évaluation.

b. 5. Informer les personnes testées avant le *testing*, sur les objectifs de l'évaluation, la nature des tests utilisés, à qui les résultats seront communiqués et l'utilisation qu'on envisage de faire de ces résultats.

b. 6. Avertir de la date à laquelle les tests seront administrés, de la date à laquelle les résultats seront disponibles, et si les personnes testées – ou

1. La législation sur cette question varie selon les pays. Par exemple, le *UK Data Protection Act* actuel donne des droits d'accès aux données archivées sur fichier électronique différents des droits d'accès aux données archivées dans des dossiers papier. NDT : En France, ces questions sont du ressort de la Commission Nationale de l'Informatique et des Libertés (CNIL). Toute collecte d'informations sur les personnes est régie par la Loi Informatique et Libertés de janvier 1971.

d'autres – pourront ou ne pourront pas, avoir une copie des tests, des feuilles de réponse auxquelles elles ont, elles-mêmes, répondu, et de leurs scores[1].

b. 7. Faire administrer les tests par une personne formée et faire interpréter les résultats par une personne qualifiée.

b. 8. S'assurer que les personnes testées sont informées du caractère facultatif éventuel d'un test et dans un tel cas, des conséquences de la passation ou non de ce test.

b. 9. S'assurer que les personnes testées comprennent les conditions, si c'est le cas, selon lesquelles elles peuvent repasser les tests, demander une vérification de la cotation des tests qu'elles ont passés, voire demander l'annulation de leurs scores.

b. 10. S'assurer que les personnes testées savent que leurs résultats leur seront expliqués aussi tôt que possible après la passation du test dans des termes facilement compréhensibles.

b. 11. S'assurer que les personnes testées comprennent que leurs résultats sont confidentiels dans les limites autorisées par la loi et les pratiques correctes.

b. 12. Informer les personnes testées de qui aura accès à leurs résultats et à quelles conditions leurs scores seront communiqués.

b. 13. S'assurer que les personnes testées sont averties des procédures pour porter plainte ou signaler un problème.

Les utilisateurs de tests informeront les personnes testées qu'on attend d'elles :

b. 14. Qu'elles traitent les autres avec courtoisie et respect pendant le processus de *testing*.

b. 15. Qu'elles posent des questions avant le début du *testing*, si elles ne sont pas sûres des raisons pour lesquelles le test est administré, de la manière dont il sera administré, de ce qu'il faudra faire et de ce qu'il adviendra des résultats.

b. 16. Qu'elles informent une personne compétente sur tout incident dont elles croient qu'il peut rendre les résultats du test invalides ou qu'elles veulent voir pris en considération.

b. 17. Qu'elles suivent les instructions de celui qui administre les tests.

---

1. Alors que les tests et les feuilles de réponses ne sont jamais communiqués aux personnes testées, il existe des différences selon les pays dans les pratiques concernant les éléments que les personnes testées – ou d'autres – peuvent obtenir. Quoiqu'il en soit, il y a davantage de différences dans les attentes des personnes testées en ce qui concerne les informations qu'on leur donnera. Il est important que le contrat clarifie ce qu'elles *n'*auront *pas* aussi bien que ce qu'elles auront.

b. 18. Qu'elles doivent être conscientes des conséquences de ne pas passer un test si elles choisissent de ne pas le passer, et être prêtes à en accepter les conséquences.

b. 19. Qu'elles s'assurent que, au cas où elles doivent payer pour la passation des tests, le paiement sera fait à la date indiquée.

➤ Annexe C : Points à prendre en considération lorsqu'on fait des aménagements pour tester des personnes présentant des infirmités ou des handicaps.

Des précautions considérables et une expertise solide sont requises lorsque le mode d'administration d'un test doit être modifié pour s'adapter aux besoins de personnes handicapées. Comme toujours, les lois nationales et locales ainsi que les pratiques[1] ont besoin d'être prises en considération, ainsi que le respect de la vie privée des individus. La demande d'informations en ce qui concerne les types et le niveau de handicap doit être limitée à la capacité de mettre en œuvre les activités demandées pour le test. Une vigilance particulière doit être exercée dans le domaine de la sélection professionnelle[2].

Il n'y a pas de principe de base simple utilisable pour s'assurer qu'un test est administré de manière équitable à des personnes quel que soit leur type de handicap. C'est une question de jugement professionnel de savoir s'il vaut mieux utiliser un type d'évaluation alternatif ou modifier le test ou ses conditions d'administration. En pratique, il est rarement possible de disposer pour des tests modifiés de normes sur des échantillons suffisants de personnes présentant des handicaps équivalents permettant d'assurer la comparabilité du test avec la version habituelle. Cependant, lorsqu'il existe des données, par exemple, sur la modification du temps accordé, l'utilisation du Braille ou de versions orales enregistrées sur bandes magnétiques des tests, de telles données devraient guider l'utilisateur pour procéder aux aménagements nécessaires. Bien qu'il ne soit pas toujours possible de réaliser une standardisation complète de la version modifiée, une étude pilote sur de petits échantillons devrait être conduite chaque fois que c'est réalisable.

1. Aux États-Unis, par exemple, on doit faire attention aux dispositions du *Americans with Disabilities Act* (1990). Au Royaume-Uni, le *Disability Discrimination Act* (1995), *Employment Code of Practice* stipule que « les employeurs sont tenus de réviser les tests – ou la manière dont les résultats à de tels tests sont évalués – pour prendre en compte les candidats présentant des infirmités spécifiques ».

2. Pour des conseils détaillés à ce sujet aux États-Unis, voir Eyde, Nestor, Heaton and Nelson (1994).

Étant donné le manque d'informations sur la performance aux tests (qu'il ait été modifié ou non) des personnes présentant un handicap, il est souvent plus approprié d'utiliser le résultat au test de manière plutôt qualitative. Ces résultats peuvent être utilisés pour fournir des indications sur les caractéristiques évaluées (aptitudes, motivation, personnalité, etc.) qui peuvent être complétées et étayées par des informations collectées en recourant à d'autres méthodes.

Pour une évaluation individuelle, l'évaluateur peut habituellement adapter les procédures d'évaluation aux possibilités de la personne évaluée. Cependant, des questions particulières se posent lors d'un *testing* collectif (par exemple, pour la sélection professionnelle). Dans ce domaine, il peut y avoir des difficultés pratiques rencontrées lorsqu'on introduit des variations dans le mode d'administration pour certains individus au sein d'un dispositif d'administration en groupe. En outre, toutes les parties peuvent considérer les différences de traitement comme inéquitables. Par exemple, si on leur accorde plus de temps pour terminer le test, ceux qui sont handicapés peuvent être conscients du fait qu'ils sont traités « différemment », et ceux qui ne sont pas handicapés peuvent avoir l'impression que ce temps supplémentaire procure un avantage inéquitable.

Des conseils sur les besoins particuliers peuvent en général être recueillis auprès des organisations de handicapés concernées, aussi bien qu'à titre individuel, auprès des personnes testées. C'est généralement utile (lorsque la loi l'autorise) de demander directement à la personne, d'une façon rassurante, si certains éléments doivent être pris en considération[1]. Dans la plupart des cas, une telle consultation permettra d'effectuer des modifications appropriées à l'environnement de passation des tests sans nécessiter de modifications du test lui-même.

L'ébauche de protocole suivante fournit un guide général pour le processus de prise de décision visant à modifier le *testing* et sur la manière de réaliser la modification[2]. Essentiellement, le handicap peut 1) ne contribuer en rien à la variance du test, 2) y contribuer de manière pertinente, ou 3) contribuer de manière non pertinente à la variance du construit mesuré. Dans le premier cas (1), aucune modification n'est nécessaire. Dans le dernier cas (3), l'objectif des modifications devrait être d'ôter la source de variance non

1. Au Royaume-Uni, *the Disability Discrimination Act* (1995) rend également obligatoire pour les individus de faire connaître leurs besoins.

2. NDT : Aucune modification ne doit être apportée à une procédure de testing sans une autorisation explicite des ayants droit.

pertinente (à l'aide de modifications appropriées de l'environnement de passation du test ou son remplacement par un test plus approprié). Dans le second cas (2) (contribution de manière pertinente à la variance du construit mesuré), quoiqu'on fasse, des modifications apportées au test auront un effet sur la pertinence des scores au test.

c1. Est-il plausible que le handicap ait un effet sur la performance au test ? De nombreuses personnes ont des handicaps qui ne devraient pas affecter la performance au test. Dans de tels cas, cela ne serait pas approprié de faire des aménagements pour elles.

c2. S'il est plausible que le handicap affecte la performance au test, alors est-ce que l'effet sur la performance est secondaire par rapport au construit mesuré ? Par exemple, une personne atteinte d'arthrose d'une main peut rencontrer des difficultés dans un test en temps limité qui fait appel à l'écriture. Si l'aptitude à réaliser des tâches manuelles rapidement fait partie intégrante du construit mesuré, alors le test ne devrait pas être modifié. Toutefois, si l'objectif de l'évaluation est la vitesse de balayage visuel, alors un mode de réponse alternatif serait approprié.

c3. Lorsqu'un handicap particulier est secondaire par rapport au construit mesuré mais peut, de manière plausible, affecter la performance individuelle au test, alors on peut envisager d'apporter des modifications à la procédure.

c4. Les utilisateurs devraient toujours consulter le manuel du test ainsi que l'éditeur pour des conseils sur les modifications et pour des informations sur les formats et les procédures alternatives envisageables.

c5. Les utilisateurs devraient également consulter les organisations de handicapés pertinentes pour des avis et des conseils sur les implications possibles d'un handicap spécifique, la littérature ou la documentation pertinente, et les types d'adaptations ou d'aménagements qui peuvent être utiles.

c6. Toute modification faite au test ou aux procédures d'administration du test devrait être soigneusement consignée et accompagnée des justifications sous-tendant cette modification.

Bibliographie

ARBISIO C. (2003), *Le bilan psychologique avec l'enfant. Approche clinique du WISC-III*. Paris, Dunod.

AUBRET J. & BLANCHARD S. (2005), *Pratique du bilan personnalisé.* Paris, Dunod.

BACHER F. (1982), Sur certains problèmes soulevés par l'utilisation des tests psychologiques, *L'Année Psychologique, 82,* 439-455.

BALICCO C. (1997), *Les méthodes d'évaluation en ressources humaines. La fin des marchands de certitude.* Paris, Editions d'Organisation.

BALICCO C. (1999), *Approche des mécanismes de prise de décision dans le choix et l'utilisation des méthodes d'évaluation et de sélection dans le recrutement des cadres en France,* Doctorat de Psychologie, Paris.

BALINSKY B. (1941), An analysis of the mental factors in various age groups from nine to sixty, *Psychologica1 Monograph, 23,* 191-234.

BALTES P.B. (1987), Theoretical propositions of life-span developmental psychology, On the dynamics between growth and decline, *Developmental Psychology, 23*, 611-626.

BALTES P.B., BALTES M.M. (1990), *Successful aging, Perspectives from the behavioral sciences.* Cambridge, Cambridge University Press.

BALTES P.B., STAUDINGER U.M. & LINDENBERGER U. (1999), Lifespan psychology, Theory and Application to Intellectual Functioning. *Annual Reviw of Psychology. 50*, 471-507.

BEAUFILS B. (1996a), *Statistiques appliquées à la psychologie. Tome 1 : statistiques descriptives.* Rosny, Bréal.

BEAUFILS B. (1996b), *Statistiques appliquées à la psychologie. Tome 2 : statistiques inférentielles.* Rosny, Bréal.

BERNAUD J-L. (2000a), *Tests et théories de l'intelligence*. Paris Dunod.

BERNAUD J-L. (2000b) Réactions au bilan psychologique le point de vue de l'usager In D. Castro (Ed.) *Les écrits en psychologie : rapports expertises bilans.* Paris l'Esprit du Temps.

BERNAUD J.-L. (2000c), Recrutement et évaluation du personnel. In J.L. Bernaud et C. Lemoine (2000). *Traité de psychologie du travail et des organisations* (pp. 95-132). Paris, Dunod.

BERNAUD J-L. (2007), *Introduction à la psychométrie.* Paris, Dunod.

BERNAUD J-L. PRIOU P. & SIMONET R. (1993), *Manuel de la NV7.* Paris, EAP.

BERNIER J-J. & PIETRULEWICZ B. (1997), *La psychométrie. Traité de mesure appliquée.* Montréal, Gaëtan Morin éditeur.

BEUSCART-ZÉPHIR M-C. & BEUSCART R. (1988), Tests de pereformance : une méthode d'analyse des startégies de résolution. Un exemple : le test de cubes du Wisc-R, *European Journal of Psychology of Education,* III-1, 33-51.

BEUSCART-ZÉPHIR M-C. & BEUSCART R. (1989), Psychologie cognitive et psychométrie : apport de l'automatisation à l'identification des processus impliqués dans les tests d'aptitudes. In J-M. Monteil et M. Fayol (Eds.), *La psychologie scientifique et ses applications.* Grenoble, Presses Universitaires de Grenoble.

BEUSCART-ZÉPHIR M-C., ANCEAUX F., DUHAMEL A. & QUENTIN S. (1996), Un exemple d'application du diagnostic cognitif, *Psychologie Française*, 41, 1, 65-76.

BINET A. (1911/1973), *Les idées modernes sur les enfants.* Paris, Flammarion.

BINET A. & SIMON T. (1905a), *La mesure du développement de l'intelligence chez les jeunes enfants*. Paris, Société A. Binet.

BINET A. & SIMON T. (1905b), Sur la nécessité d'établir un diagnostic scientifique des états inférieurs de l'intelligence, *L'Année Psychologique, onzième année, 163-244* [Article réédité en 2004. Paris, L'Harmattan].

BLANCHARD S. (2002), De l'examen d'orientation professionnelle au bilan de compétences. Actes du colloque *La place de l'évaluation dans le processus d'orientation professionnelle des adultes.* Lille, INOIP AFPA. 11-23.

BLANCHARD S. (2007), L'évaluation dans le cadre du conseil en orientation : l'exemple de la démarche de bilan de compétences, *Les Dossiers des Sciences de l'Education, 18,* 61-70.

BLANCHARD S., SONTAG J-C. & LESKOW S. (1999), L'utilisation d'épreuves conatives dans le cadre du bilan de compétences. *L'Orientation Scolaire et Professionnelle, 28, 2,* 275-297.

BONNARDEL R. (1953), Le test B 101, *Le travail Humain, 3-4,* 253-266.

BORKOWSKI J.G. & CAVANAUGH J. (1979), Maintenance and generalisation of skills and strategies by the retarded. In N. Ellis (Ed.), *Handbook of mental deficiency* (pp. 569-617). Hillsdale, NJ, Erlbaum.

BOTWINICK J. (1977), Intellectual Abilities. In J.E. Birren & K.W Schaie (Eds), *Handbook of the psychology of aging*. New york, Van Nostrand Reinhold.

BOURGES S. (1979), *Approche génétique et psychanalytique de l'enfant.* Tome 1. Neuchâtel, Delachaux et Niestle.

BOURGUIGNON O. (2000), Introduction au n° spécial du Bulletin de Psychologie, 2000, 53 (1) « Ethique en psychologie et déontologie des psychologues ».

BOURGUIGNON O. (2003), *Questions éthiques en Psychologie*. Paris, Mardaga.

BROWN A. & FERRARA R.A. (1985), Diagnosing zones of proximal development. In J. Wertsch (Ed.), *Culture, communication, and cognition, Vygotskian perspectives*, (pp.273-305), Cambridge, MA, Cambridge University Press.

BROWN A. & FRENCH L.A. (1979), The zone of potential development, Implication for intelligence testing in the year 2000. In R.J. Sternberg & D.K. Detterman (Eds.) *Human Intelligence*, pp. 217-235. Norwood, N.J., Ablex.

BRUCHON-SCHWEITZER M.L. & FERRIEUX D. (1991), Une enquête sur le recrutement en France. *Revue Européenne de Psychologie Appliquée, 41,* 1, 9-17.

BRUCHON-SCHWEITZER M.L. & LIEVENS S. (1991), Le recrutement en Europe Recherches et pratiques. *Psychologie et Psychométrie* 12,7-71.

BÜCHEL F. & PAOUR J.L. (Eds.) (1990), Assessment of learning and development potential, Theory and practices. *European Journal of Psychology of Education. 5*(2), 89-95.

BÜCHEL F., DE RIBAUPIERRE A. & SCHARNHORST U. (1990), Le diagnostic du potentiel d'apprentissage par le LPAD, une étude de la fidélité. *European Journal of Psychology of Education, 5*, 135-158.

BÜCHEL F.P. (ED.) (1995), *L'éducation cognitive, le développement de la capacité d'apprentissage et son évaluation*. Lausanne, Delachaux et Niestlé.

BUDOFF M. & CORMAN L. (1974), Demographic and psychométric factors related to improved performance on the Kohs learning-potential procedure. *American Journal of Mental Deficiency, 78*(5), 578-585.

BUDOFF M. & HAMILTON J.L. (1976), Optimizing test performance of moderately and severely-mentally retarded adolescents and adults. *American Journal of Mental Deficiency, 81*, 49-57.

BUDOFF M. (1968), Learning potential as a supplemantary testing procedure. In J. Hellmuth (Evaluation dynamique.), *Learning disorders* (vol. 3, pp.295-343). Seattle, Special Child.

BUDOFF M. (1987), Measures for assessing learning potential. In C.S. Lidz (Ed.) *Dynamic assessment,* (pp. 173-195). New York, The Guilford Press.

BUTT D.S. & BEISER M. (1987), Successful aging, a theme for international psychology. *Psychology and Aging, 2*, 87-94.

CAMPIONE J.C. & BROWN A.L. (1987), Linking dynamic assessment with school achievement. In C.S. Lidz (Ed.) *Dynamic assessment,* (pp. 82-115). New York, The Guilford Press.

CAROFF X. (2004), L'identification des enfants à haut potentiel : quelles perspectives pour l'approche psychométrique ?, *Psychologie Française, 49, 3,* 233-251.

CARPENTER P.A., JUST M.A. & SHELL P. (1990), What one intelligence test measures : A theorical account of the processing in the Raven Progressive Matrice test. *Psychological Review, 97,* 404-431.

CARROLL J.B. (1962), The prediction of success in intensive foreign language training. In R. Glaser (Ed.), *Training research and education* (pp.87-136). Pittsburgh, University of Pittsburgh Press.,.

CARROLL J.B. (1989), The Carroll model, A twenty-five year retrospective and prospective view. *Educational Researcher, 18 (1),* 26-31.

CARROLL J.B. (1993), *Human cognitive abilities, A survey of factor-analytical studies.* New York, Cambridge University Press.

CASTRO D. (2001), L'examen psychologique au moyen des tests : de la pratique professionnelle à la formation universitaire, *Le journal des psychologues, 186,* 52-55.

CASTRO D. (Ed.), (2000), *Les écrits en psychologie : rapports, expertises, bilans.* Paris, l'Esprit du Temps.

CASTRO D. (2006), *Pratique de l'examen psychologique en clinique adulte*. Paris, Dunod.

CASTRO D., MOGENET J-L., POZZI B., GLATZ N., CARDOSO C., THIEBAULT P. & PEINTURE S. (2001) Qui doit utiliser les tests psychologiques : psychologues ou non psychologues ?, *Pratiques Psychologiques, 2,* 103-118.

CASTRO D., BERNAUD J-L. (1996), Quel avenir pour les tests psychologiques au XXI[e] siècle, *Pratiques Psychologiques, 4,* 1-3.

CASTRO D., MELJAC C., JOUBERT B. (1996), Pratiques et outils des psychologues cliniciens français. Les enseignements d'une enquête, *Pratiques Psychologiques, 4,* 73-80.

CATTEL R.B. (1971), *Abilities, their structure, growth and action*, Boston, Houghton Mifflin.

CHARTIER D. (2002), La place de l'évaluation dans le processus d'orientation de l'AFPA. *Actes du colloque « La place de l'évaluation dans le processus d'orientation professionnelle des adultes »*. Lille, INOIP.

CHARTIER P. (1999), Étude de la variabilité intra et inter-individuelle dans la résolution d'une épreuve du type cubes de Kohs. In M. Huteau & J. Lautrey (Eds), *Approches différentielles en Psychologie*. 145-163. Rennes, P.U.R.

CHARTIER P. (2001), Les apports de la recherche en psychologie aux tests d'intelligence : quelles conséquences pour la pratique ? *L'Orientation Scolaire et Professionnelle, 30,* 4, 509-531.

CHARTIER P. (2002a), L'épreuve de Kohs : validité, adaptations et utilisations, *Psychologie et Psychométrie, 23, 3-4,* 21-50.

CHARTIER P. (2002b), Vers une évaluation de type « diagnostic cognitif », Actes du colloque de l'AFPA « La place de l'évaluation dans le processus d'orientation professionnelle des adultes », p.177-181. AFPA.

CHARTIER P. (2002c), *Variabilité des situations et variabilité des stratégies de résolution ? L'exemple d'une épreuve de type cubes de Kohs.* Thèse de doctorat. Université René Descartes Paris V.

CHARTIER P. (2005), Piéron et la docimologie. Quelques recherches de Piéron, et du Service de Recherche de l'INETOP, sur l'analyse de la fiabilité de l'évaluation scolaire. *L'Orientation Scolaire et Professionnelle*, 3, hors série, 257-263.

CHARTIER P. (2008b), Expérimentation d'une épreuve de facteur g utilisant comme support des cartes à jouer, In E. Loarer, P. Vrignaud, J-L. Mogenet, F. Cuisinier, H. Gottesdiener et P. Mallet, *Perspectives différentielles en psychologie, 39-42.* Rennes, Presses Universitaires de Rennes.

CHARTIER P. (à paraître, 2008a), Les tests dominos (D70 et D2000) : comment dépasser le constat du seul score total ? Exemples d'analyses des réponses, *Pratiques Psychologiques.*

CHARTIER D. & LOARER E. (1994), Évaluation dynamique de l'intelligence non-verbale par la procédure aide au cours du test, application à une population non-francophone et à des adultes de bas niveau de qualification. In M. Huteau (Ed.), *Les techniques psychologiques d'évaluation des personnes*. (pp.141-144), Issy les Moulineaux, EAP.

CHI M.T.H. (1978), Knowledge structure and memory development. In R. Siegler (Ed.), *Children's thinking, What develop ?* Hillsdale, NJ, Erlbaum, pp. 73-96.

CIANCIOLO A. & STERNBERG R.J. (2004), *Intelligence, A brief history*. Blackwell Publishing.

COGNET G. (2005), *NEMI-2, les options d'une révision.* Communication au colloque international *Intelligence de l'enfant,* Fédération Française des Psychologues et de psychologie, Paris, 6-8 Octobre.

COGNET G. (2006), Les psychologues scolaires, In J. Lautrey, *Psychologie du développement et psychologie différentielle.* p.457-470. Paris, PUF. (Collection dirigée par S. Inonescu et A. Blanchet).

COOK M. (1988), *Personnel selection and productivity,* Chichester.

CORROYER D., WOLFF M. (2003), *L'analyse statistique des données en psychologies*. Paris, Armand Colin.

CRAIK F.I., BYRD M. & SWANSON J.M. (1987), Patterns of memory loss in three elderly samples. *Psychology and Aging, 2,* 79-86.

DANY F. & TORCHY V. (1994), Recruitment and selection in Europe, policies, practices, and methods. ln C. Brewster & A. Hegewisch (Eds.). *Policy and practice in European Human Resource Management.* Routledge, London.

DAS J.P. (1987), Introduction. In C.S. Lidz (Ed.), *Dynamic assessment* (pp. Vii-xi). New-York, Guilford Press.

DE RIBAUPIERRE A. (1995), Potentiel d'apprentissage et contraintes structurales, Apports des modèles piagétiens et néo-piagétiens. In F. Büchel (Ed.) *L'éducation cognitive. Le développement de la capacité d'apprentissage et son évaluation*. (pp. 135-161). Neuchâtel, Delachaux et Niestlé.

DE SHON R-P., CHAN D. & WEISSBEIN D.A. (1995), Verbal overshadowing effects on Raven's Advanced Progressive Matrices : evidence for multidimensional performance determinants, *Intelligence,* 21, 135-155.

DEVOUCHE E. (2003), Les banques d'items. Construction d'une banque pour le Test de Connaissance du Français, *Psychologie et Psychométrie, 24, 2/3,* 89-116.

DICKES P. (1988), Configurations perceptives et difficulté des stimuli construits d'après la technique de Kohs, *Bulletin de Psychologie, XLII, 388,* 210-218.

DICKES P. (1999), Modèles de réponse à l'item (MRI) et recherche en psychologie, *Psychologie et Psychométrie, 20, 2/3,* 8-18.

DICKES P., HOUSSEMAND C. & REUTER M. (1996), Modèles pour le contenu des tâches d'assemblage de faces géométriques et difficulté des items. *Psychologie Française, 41,1*, 47-55.

DICKES P., MARTIN R. (1998), Les composantes de l'intelligence générale du D70. *Psychologie et Psychométrie*, *19, 1*, 27-51.

DICKES P., TOURNOIS J., FLIELLER A. & KOP J.L. (1994), *La psychométrie*, Paris, PUF.

EAP (1978), *Manuel des cubes de Kohs.* Paris, EAP.

ECKERT P., LUDWIG C. & RAFFIN D. (2008), Table ronde franco-allemande sur les méthodes et outils du bilan de compétences. Communication au colloque « Autour des compétences », Université de Rouen, 22 mai 2008.

ECPA (1961), *Test D48. Manuel d'application*. Paris, ECPA.

ECPA (1970), *Test D70. Manuel d'application*. Paris, ECPA.

ECPA (2000a), *Test D2000. Manuel d'application*. Paris, ECPA.

ECPA (2000b), *Test R2000. Manuel d'application*. Paris, ECPA.

EMBRETSON S. E. (1987), Toward development of a psychometric approach. In C.S. Lidz (Ed.), *Dynamic assessment* (pp. 141-170). New-York, Guilford Press.

EMBRETSON S.E. (1989), Latent trait models as an information-processing approach to testing. *International Journal of Educational Research, 13*, 189-203.

EMBRETSON S.E. (1991), A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56* (3), 495-515.

EMBRETSON S.E. (1995), A measurement model for linking individual learning to processes and knowledge, Application to mathematical reasoning. *Journal of Educational Measurement, 32*, 277-294.

FAVERGE J.M. (1955), *Calcul des longueurs, test*. Braine-le-Château, Applications des techniques modernes.

FEUERSTEIN R. (1980), *Instrumental Enrichment*. Baltimore, University Park Press.

FEUERSTEIN R. (1990), Le PEI. In J. Martin & G. Paravy (Eds.), *Pédagogies de la médiation*. Lyon, Chroniques sociales.

FEUERSTEIN P., HOFFMAN M.B., JENSEN M.R. &, RAND Y. (1985), Instrumental enrichment, an intervention program for structural cognitive modifiability, theory and practice, In J.W. Segal, S.F. Chipman & R. Glaser (Eds.), *Thinking and learning skills*, Vol. 1. Hillsdale, NJ, Erlbaum.

FEUERSTEIN R., RAND Y., HOFFMAN M. & MILLER R. (1979), Cognitive modifiability in retarded adolescents. Effects of Instrumental Enrichment. *American Journal for Mental Deficiency, 83*, 539-550.

FEUERSTEIN R., RAND Y. & HOFFMAN M.B. (1979), *The dynamic assessment of retarded performers, the learning potential assessment device, theory, instruments, and techniques*. Glenview, IL, Scott, Foresman & Co.

FEUERSTEIN R., RAND Y., HOFFMAN M. & MILLER R. (1980), *Instrumental enrichment, an intervention program for cognitive modifiability*. Baltimore, MD, University Park Press.

FEUERSTEIN R., RAND Y., JENSEN M.R. KANIEL S. & TZURIEL D. (1987), Prerequisites for assessing of learning potential, the LPAD model. In C.S. Lidz (Ed.) *Dynamic assessment* (pp. 35-51). New York, The Guilford Press.

FLAMMER A. & SCHMID H. (1982/1995), Tests d'apprentissage, concept, réalisation, évaluation. In F. Büchel (Ed.) *L'éducation cognitive. Le développement de la capacité d'apprentissage et son évaluation*. (pp. 179-214). Neuchâtel, Delachaux et Niestlé.

FLIELLER A., (1999), Etude d'un texte lexical (définitions lacunaires) par des modèles de réponse à l'item, *Psychologie et Psychométrie, 20, 2/3,* 65-84.

FLIELLER A. (2001), Problèmes et stratégies dans l'explication de l'effet Flynn. In M. Huteau (Ed.), *Les figures de l'intelligence*. Paris, EAP.

FLIELLER A., (2002), Quelques remarques sur la mesure en psychologie, *Bulletin de Psychologie, 55-6,* 555-560.

FLIELLER A., SAINTIGNY N. & SCHAEFFER R. (1986), L'évolution du niveau intellectuel des enfants de 8 ans sur une période de 40 ans (1944-1984). *L'Orientation Scolaire et Professionnelle*, 15, 61-83.

FLYNN J.R. (1984), The mean IQ of Americans, massive gains 1932 to 1978, *Psychological Bulletin*, 95, 29-51.

FLYNN J.R. (1987), Massive gains in 14 nations, what IQ tests really *measure, Psychological Bulletin, 101*, 171-191.

FONTAINE R. (1999), *Manuel de Psychologie du Vieillissement.* Paris, Dunod.

FRAISE J-P. (1991), La psychométrie à l'AFPA : son rôle dans l'orientation et le recrutement des demandeurs de formation professionnelle, *L'Orientation Scolaire et Professionnelle*, 20 (1), 127-139.

GARDNER H. (1996), *Les intelligences multiples*. Paris, Retz.

GARDNER H. (1999), *Les formes de l'intelligence* (1e éd. 1983*, Frames of mind, the theory of multiple intelligences*). Paris, Odile Jacob.

GAUDRON J-P. (1999), La psychométrie assistée par ordinateur : problématiques en question et perspectives de recherches, *L'Orientation Scolaire et Professionnelle, 28, 1,* 31-62.

GAUDRON J-P. (2008), *Internet, diagnostic informatisé et bilan de compétences.* Communication au colloque « Autour des compétences », Université de Rouen, 22 mai 2008.

GAVAND A. (2006), *Prévenir la discrimination à l'embauche*. Paris, Editions d'Organisation.

GILLES P-Y. (1991), Etude des différences individuelles dans les stratégies de résolution d'une épreuve de visualisation spatiale, *Actes des IXes journées de psychologie différentielle, 188-20.* Liège, Presses Universitaires de Liège.

GILLES P-Y. (1993), *Etude des différences individuelles dans les stratégies de résolution de problèmes spatiaux.* Thèse de doctorat. Université René Descartes Paris V.

GILLET B. (1987). Aptitudes et capacités cognitives. In C Lévy-Leboyer et C. Spérandio (Eds.) *Traité de Psychologie du Travail*. Paris, PUF.

GLASER R. & PELLEGRINO (1982), Improving the skills of learning. In D.K. Detterman & R.J. Sternberg (Eds.), *How and how much can intelligence be increased*, pp. 197-212. Norwood, N.J., Ablex.

GOLDSTEIN K. & SCHEERER M. (1941), The Goldstein-Scheerer cube test, *Psychological Monographs, 35, 2,* 32-56.

GOULD S.J. (1983), *La mal-mesure de l'homme*. Paris, Ramsay.

GRÉGOIRE J. (1996), Grille d'aide à l'interprétation des scores aux différents subtests du WISC-III Paris, ECPA.

GRÉGOIRE J. (2000a), *L'évaluation clinique de l'intelligence de l'enfant. Théorie et pratique du WISC-III.* Liège, Mardaga.

GRÉGOIRE J. (2000b), Orientation, évaluation et facteurs culturels, in AFPA, *Actes du colloque « La place de l'évaluation dans le processus d'orientation professionnelle des adultes »,* p.103-112. Montreuil, AFPA.

GRÉGOIRE J. (2004), *L'examen clinique de l'intelligence de l'adulte*. Liège, Mardaga.

GRÉGOIRE J. (2005), Les métamorphoses des échelles de Wechsler, *Questions d'orientation, 4,* 53-59.

GRÉGOIRE J. (2006), *L'examen clinique de l'intelligence de l'enfant. Fondements et pratique du WISC-IV.* Sprimont, Mardaga.

GRIGORENKO E.L. & STERNBERG R.J. (1998), Dynamic testing. *Psychological Bulletin, 124* (1), 75-111.

GUÉDON M-C. & SAVARD R. (2000), *Tests à l'appui. Pour une intervention intégrée de la psychométrie en counseling d'orientation.* Québec, les Editions Septembre.

GUÉGUEN N. (2005), *Statistiques pour psychologues.* Paris, Dunod.

GUICHARD J., HUTEAU M. (2006, 2e édition), *Psychologie de l'orientation.* Paris, Dunod.

GUILLEVIC C., VAUTIER S. (1998), *Diagnostic et tests psychologiques.* Paris, Nathan.

GUSTAFFSON J.E. (1984), An unifying model for the structure of intellectual abilities. *Intelligence*, 8, 179-203.

GUSTAFFSON J.E. (1988), Hierarchical models of individual differences in cognitive abilities. In R.J. Sternberg (Ed.) *Advances in the psychology of human intelligence (Vol.4)* Hillsdale, NJ, Erlbaum.

GUTHKE J. (1982), The learning test concept, an alternative to the traditional static intelligence test. *The German Journal of Psychology, 6* (4), 306-324.

GUTHKE J. (1990), Les tests d'apprentissage comme alternative ou complément aux tests d'intelligence, un bilan de leur évolution. *European Journal of Psychology of Education, 5* (2), 117-133.

GUTHKE J. (1992), Learning tests, The concept, main research findings, problems and trends. *Learning and Individual Differences, 4*, 137-151.

GUTHKE, J. & BECKMANN J.F (2000), The learning test concept and itt application in practice. In C.S. Litz & J.G. Elliot (Eds.), *Dynamic assessment: Prevailing models and applications* (pp.17-69). Oxford, England, Elsevier.

GUTHKE J. & BECKMANN J.F. & DOBAT H. (1997), Dynamic testing, problems, uses, trends and evidence of validity. *Educational and Child Psychology, 14* (4) 17-32.

GUTHKE J. & WINGENFELD S. (1992), The learning test concept, Origin, state of the art, and trends. In H.C. Haywood & D. Tzuriel (Eds.) *Interactive Assessment*, pp. 64-94. New York, Springer.

GUTTMAN L. (1957), Empirical Verification of the Radex Structure of Mental Abilities and Personality Trait, *Educational and Psychological Measurement*, 17, 391-407.

GUTTMAN L. (1965), The Structure of interrelations among intelligence tests. Invitational Conference on Testing Problems. Princeton, New Jersey, Educational Testing Service, 25-36.

GUTTMAN L. & LÉVY S. (1991), Two structural laws for intelligence tests. *Intelligence*, 15, 79-103.

HAMBLETON R.K. & SLATER S. (1997), Item response theory models and testing practices, current international status and futur directions. *European Journal of Psychological Assessment, 13*(1), 21-28.

HAMBLETON R.K., SWAMINATHAN H. & ROGERS H.J. (1991), *Fundamentals of item response theory*, Newbury Park, Ca, Sage.

HARTIGAN J.A. & WIGDOR A.K. (1989), *Fairness in employment testing, Validity generalization, minority issues and the General Aptitude Test Battery*. Washington, DC, National Academy Press.

HAYWOOD H.C. & TZURIEL D. (Eds.) (1992), *Interactive testing*. New York, Springer Verlag.

HERTZOG C., SCHAIE K.W. & GRIBBIN K. (1978), Cardiovascular disease and changes in intellectual functioning from middle to old age. *Journal of Gerontology*, *33,* 872-883.

HORN J.L. (1970), Organization of data on life-span development of human abilities. In L.R. Goulet et P. B. Baltes (Eds.) *Life-span developmental psychology*. New york, Academic Press.

HORN J. L. & CATTELL R.B. (1966), Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology*, 57, 253-270.

HOUSSEMAND C. (1999a), *Adaptabilité stratégitaire dans la résolution des cubes de Kohs.* Thèse de doctorat, Université Nancy 2.

HOUSSEMAND C. (1999b), Approche méthodologique pour l'étude et l'analyse des stratégies de résolution. L'exemple des cubes de Kohs. In M. Huteau & J. Lautrey (Eds), *Approches différentielles en Psychologie*. 213-218. Rennes, P.U.R.

HUNT E. (1974), Quote the Raven ? Nevermore ! In J. Gregg (Ed.), *Knowledge and Cognition,* Hillsdale N.J., Erlbaum, 129-157.

HUNTER J.E., HUNTER R.F. (1984), Validity and utility of alternate predictors,of job performance. *Psychological Bulletin*, 96, 72-98.

HUNTER J.E., SCHMIDT F.L. (1996), Intelligence and job performance, economic and social implications. Psychology, Public Policy, and Law, 2, 447472.

HURTIG M. (I995), Constat d'acquisition ou pronostic d'apprentissage. Peut-on dynamiser la psychométrie ? In F. Büchel (Ed.) *L'éducation cognitive. Le développement de la capacité d'apprentissage et son évaluation*. (pp. 165-178). Neuchâtel, Delachaux et Niestlé.

HUTEAU M. (1985), Dimensions des différences individuelles dans le domaine intellectuel et processus de traitement de l'information. In J. Drevillon, M. Huteau, F. Longeot, M. Moscato et T. Ohlmann, *Fonctionnement cognitif et individualité,* 41-87. Bruxelles, Pierre Mardaga.

HUTEAU M. (1994), L'évaluation psychologique des personnes : problèmes et enjeux actuels, *in* M. Huteau (éd.), *Les techniques psychologiques d'évaluation des personnes.* Issy-les-Monlineaux, EAP.

HUTEAU M. (1996), L'évaluation par les notes et par les tests. In Lieury, *Manuel de psychologie de l'éducation et de la formation*. Paris, Dunod.

HUTEAU M. (2002, 2e éd.), *Psychologie différentielle. Cours et exercices*. Paris, Dunod.

HUTEAU M. (2005), *Écriture et personnalité. Approche critique de la graphologie*. Paris, Dunod.

HUTEAU M. (2006) Les conseillers d'orientation, In J. Lautrey, *Psychologie du développement et psychologie différentielle.* pp. 483-494. Paris, PUF. (Collection dirigée par S. Inonescu et A. Blanchet).

HUTEAU M. (Ed.). (2001), *Les figures de l'intelligence*. Paris, EAP.

HUTEAU M. & LAUTREY J. (1975), Artefact et réalité dans la mesure de l'intelligence (à propos du livre de Michel Tort, le QI). *L'Orientation Scolaire et Professionnelle*, 4, 169-187.

HUTEAU M. & LAUTREY J. (1997), *Les tests d'intelligence*. Paris, La découverte.

HUTEAU M. & LAUTREY J. (1999a), *Evaluer l'intelligence. Psychométrie cognitive*. Paris, PUF.

HUTEAU M. & LAUTREY J. (1999b) (Eds), *Approches différentielles en Psychologie*. Rennes, P.U.R.

HUTEAU M. & LAUTREY J. (2006), *Les tests d'intelligence*. Paris, Editions La Découverte.

HUTEAU M. & LOARER E. (1992), Comment évaluer les méthodes d'éducabilité cognitive ? *L'Orientation Scolaire et Professionnelle, 21* (1), 47-74.

IONESCU S. & JOURDAN-IONESCU C. (1984/85), L'évaluation du potentiel d'apprentissage 1- Utilisation du test des cubes. *Bulletin de Psychologie*, *38* (372), 919-927.

IONESCU S., JOURDAN-IONESCU C., ALAIN S. (1986/87), L'évaluation du potentiel d'apprentissage 2- Une nouvelle méthode de quantification. *Bulletin de Psychologie*, *40* (380), 481-487.

JENSEN M.R., FEUERSTEIN R. (1987), The learning potential assessment device, from philosophy to practice. In C.S. Lidz (Ed.) *Dynamic assessment* (pp. 379-402*)*, New York, The Guilford Press.

JONES H.E. & CONRAD H.S. (1933), The growth and decline of intelligence, A study of a homogeneous group between the ages of ten and sixty. *Genetic Psychology Monographs, 13*, 223-298.a.

JUHEL J. (1999), Coordination du numéro spécial sur les modèles MRI, *Psychologie et Psychométrie, 20, 2/3,* 8-18.

JUHEL J. (2005), La psychométrie: la recherche et l'évaluation des compétences qui caractérisent l'intelligence. In J. Lautrey et J.F. Richard (Dir.), *L'intelligence. Traité des Sciences cognitives, pp.23-39.* Paris, Hermès.

KAUFMAN A. (2001), Tendances actuelles dans le domaine de l'évaluation de l'intelligence, *Psychologie Française, 46, 3,* 271-280.

KAUFMAN A. (2005), *Capacité d'apprentissage, capacité de planification et le nouveau KABC-II.* Communication au colloque international *Intelligence de l'enfant,* Fédération Française des Psychologues et de psychologie, Paris, 6-8 Octobre.

KAUFMAN A. & KAUFMAN N. (1993), *Manuel du K-ABC.* Paris, ECPA.

KLAUER K.J. (1990), A process theory of inductive reasoning tested by the teaching of domain-specific thinking strategies. *European Journal of Psychology of Education. 5* (2), 191-206.

KLAUER K.J. (1995), Les effets d'entraînement de la pensée sont-ils généraux ou spécifiques ? Un apport à la vérification de la théorie prescriptive de la pensée inductive. In F.P. Büchel (Ed.), *L'éducation cognitive, le développement de la capacité d'apprentissage et son évaluation*. (pp. 285-305). Lausanne, Delachaux et Niestlé.

KREUTZ M. (1934), Comment remédier à l'inconstance des tests. *Archives de Psychologie,* 227-244.

LABERON S., LAGABRIELLE C. & VONTHRON A.M. (2005), Examen des méthodes d'évaluation dans les pratiques de sélection et d'orientation professionnelles, le cas du recrutement et du bilan de compétences. *Revue Internationale de Psychologie du Travail et des Organisations,* l (11), 3-14.

LARCEBEAU S. (1967). Évolution de la structure factorielle des aptitudes au cours du cycle d'observation, *BINOP, 23*, 261-277.

LAROCHE J.L. (1956), L'analyse des erreurs sur le Matrix 38. *Bulletin du Centre d'Etudes et Recherches Psychotechniques, 6*(2), 161-174.

LATHOUD S. (1997), L'examen d'expertise auprès des commissions de l'éducation spéciale, In Guillard et Guillemard (Eds.), *Manuel pratique de psychologie en milieu éducatif*. Paris, Masson.

LAUGHTON P. (1990), The dynamic Assessment of intelligence, a review of three approaches. *School Psychology Review, 19*(4), 459-470.

LAUTREY J. (1994), L'évaluation du potentiel d'apprentissage, état de la question. In M. Huteau (Ed.) *Les techniques psychologiques d'évaluation des personnes*. (pp.134-140), Issy les Moulineaux, EAP.

LAUTREY J. (1995), Les apports de la psychologie cognitive à la compréhension des différences en matière d'intelligence et de réussite scolaire, in Blanchet. G et al., *Intelligences, scolarité et réussite*. Paris, La pensée sauvage.

LAUTREY J. (1999), Histoire et évolution de la psychologie différentielle. In P-Y. Gilles (Ed.), *Psychologie Différentielle.* Rosny, Bréal.

LAUTREY J. (2001), L'évaluation de l'intelligence : état actuel et tentatives de renouvellement. In M. Huteau, (Ed.), *Les figures de l'intelligence*. Paris, EAP.

LAUTREY J. (2004), Etat de la recherche sur la précocité intellectuelle, *Psychologie Française, 49, 3,* 219-352.

LAUTREY J. (2007), Pour l'abandon du QI, les raisons du succès d'un concept dépassé. I*n* M. Duru-Bellat et M. Fournier (Eds.), *L'Intelligence de l'enfant. L'empreinte du social.* Sciences Humaines Éditions.

LAUTREY J., DE RIBAUPIERRE A. & RIEBEN L. (1986), Les différences dans la forme du développement cognitif évalué avec des épreuves piagétiennes, une application de l'analyse des correspondances. *Cahiers de Psychologie Cognitive, 6*, 575-613.

LAUTREY J. & RICHARD J.F. (2005), *L'intelligence. Traité des Sciences cognitives*. Paris, Hermès, 17-20.

LAVEAULT D. & GRÉGOIRE J. (1997), *Introduction aux théories des tests en sciences humaines*. Paris, De Boeck.

LAVEAULT D. & GRÉGOIRE J. (2002), *Introduction aux théories des tests en sciences humaines*. Bruxelles, De Boeck Université.

LAVEAULT D. & GRÉGOIRE J. (2002), *Introduction aux théories des tests*. Paris, De Boeck.

LEMAIRE P. & BEHRER L. (2005), *Psychologie du vieillissement, une perspective cognitive*. Bruxelles, De Boeck.

LÉVY-LEBOYER C. (1987), Problèmes éthiques posés par l'usage des tests, In C., Lévy-Leboyer et J-C., Spérendio (Eds.), *Traité de psychologie du travail, 473-485*. Paris, PUF.

LÉVY-LEBOYER C. (1990), *Evaluation du personnel. Quelles méthodes choisir ?* Paris, Les éditions d'Organisation.

LÉVY-LEBOYER C. (1996), *Évaluation du personnel, Quels objectifs ? Quelles méthodes ?* Paris, Eyrolles.

LÉVY-LEBOYER C. (2002), *Évaluation du personnel, Quels objectifs ? Quelles méthodes ?* Paris, Editions d'organisation.

LHOTTELIER A. (2000), L'acte de tenir conseil. *L'Orientation Scolaire et Professionnelle, 29, 1,* 27-50.

LIDZ C.S. (1991), *Practitioner's guide to dynamic testing*. New York, Guilford Press.

LIDZ C.S. (Ed.) (1987), *Dynamic assessment, an interactional approach to evaluating learning potential*. New York, The Guilford Press.

LIDZ C.S. & THOMAS C. (1987), The preschool learning assessment device, extention of a static approach. In C.S. Lidz (Ed.) *Dynamic assessment* (pp. 288-326*)*. New York, The Guilford Press.

LINDENBERGER U. & BALTES P.B. (1997), Intellectual functioning in old and very old age, cross-sectional results from the Berlin Aging Study. *Psychology and Aging, 12*(3), 410-432.

LOARER E. (1998), L'éducation cognitive, modèles et méthodes pour apprendre à penser. *Revue Française de Pédagogie, 122* (1), 121-161.

LOARER E. (2000), *La plasticité cognitive. Apprentissage et développement cognitif chez l'adulte, une approche différentielle*. Synthèse d'Habilitation à Diriger des Recherches. Paris, Université !rené Descartes.

LOARER E. (2001), L'évaluation dynamique comme moyen de limiter les biais culturels dans les tests d'intelligence. In M. Huteau (Ed.), *Les figures de l'intelligence*. Paris, EAP.

LOARER E. (2005), L'intelligence sociale et l'intelligence émotionnelle. In J. Lautrey et J.F. Richard (Dir.), *L'intelligence. Traité des Sciences cognitives*. Paris, Hermès, 91-107.

LOARER E. & CHARTIER D. (1994), Le potentiel d'apprentissage est-il général ou spécifique au domaine d'apprentissage ? In M. Huteau (Ed.) *Actes du Colloque international "Les techniques psychologiques d'évaluation des personnes"*. Paris (25-27 mai 1993), (pp. 150-154). Paris, EAP.

LOARER E. & CHARTIER D. (1996a), L'évaluation dynamique des aptitudes, révolution ou gadget ? In J.F. Richard (Ed.) Numéro spécial sur le diagnostic cognitif. *Psychologie Française, 41*(1), 35-46.

LOARER E. & CHARTIER D. (1996b), *Etude de validation d'une épreuve de potentiel d'apprentissage utilisable pour pronostiquer l'adaptation scolaire des jeunes étrangers nouvellement arrivés en France.* Rapport au Ministère de l'Education Nationale, Rectorat de Paris et au Fonds d'Action Sociale (F.A.S.) (24 p.).

LOARER E., CHARTIER D., HUTEAU M. & LAUTREY J. (1995a), *Peut-on éduquer l'intelligence ? L'évaluation d'une méthode d'éducation cognitive*. Berne, Peter Lang.

LOARER E., LAUTREY J., HUTEAU M. & CHARTIER D. (1995b), Effets d'une méthode de remédiation cognitive sur une population d'adultes faiblement qualifiés. *Enfance*, *2,* 263-271.

LOARER, E., VRIGNAUD P., MOGENET J-L., CUISINIER F., GOTTESDIENER H. & MALLET P. (2008), *Perspectives différentielles en psychologie.* Rennes, Presses Universitaires de Rennes.

LOHMAN D.F. (2000), Complex information processing and intelligence. In R.J. Sternberg (Ed.). *Handbook of intelligence*, pp. 285-340. Cambridge, Cambridge University Press.

LUBART T. (Ed.) (2006), *Enfants exceptionnels. Précocité intellectuelle, haut potentiel et talents.* Rosny-sous-Bois, Bréal.

MARQUER J. & PEREIRA M. (1987), Evolution à long terme des stratégies dans une tâche de comparaison phrase-dessin, *L'Année Psychologique, 87,* 329-343.

MARQUER J. & PEREIRA M. (1990), Les stratégies dans la vérification phrase-dessin. In M. Reuchlin, F. Longeot, C. Marendaz et T. Ohlmann (Eds.), *Connaître différemment.* Nancy, Presses Universitaires de Nancy..

MARQUIÉ J.-C. (1997), Vieillissement cognitif et expérience, l'hypothèse de la préservation, *Psychologie Française, 42*(4), 333-344.

MARTIN R. (2003), Le testing adaptatif par ordinateur dans la mesure en éducation : potentialités et limites, *Psychologie et Psychométrie, 24, 2/3,* 89-116.

Mc GHEE R. (1993), Fluid and crystallized intelligence, Confirmatory factor analyses of the Differential Ability Scales, Detroit Tests of Learning Aptitude-3, and Woodcock-Johnson Psycho-Educational Battery-Revised. *Journal of Psychoeducational Assessment.*

MILES C.C. & MILES W.R. (1932), The correlation of intelligence scores and chronological age from early to late maturity. American Journal of Psychology, 44, 44-78.

NAGLIERI J.A. (1998), *Manuel du NNAT.* Paris, ECPA.

NELSON E.A. & ANNEFER D. (1992), Aged heterogeneity, fact or fiction ? The fate of diversity in gerontological research. *The Gerontologist, 32*, 17-23.

NESSELROADE J.R. & THOMPSON W.W. (1995), Selection and related threats to group comparisons, an example comparing factorial structures of higher and lower ability groups of adult twins. *Psychological Bulletin* 117, 271-84.

NGUYEN-XUAN A. (1969). Etude par le modèle factoriel d'une hypothèse sur les processus de développement, *BINOP, 25.*

NOIZET G. & CAVERNI J-P. (1978), *Psychologie de l'évaluation scolaire.* Paris, PUF.

OHLMANN T. (1990a), Les systèmes perceptifs vicariants. In M. Reuchlin, J. Lautrey, C. Marendaz et T. Ohlmann, *Cognition : l'individuel et l'universel.* Paris, PUF.

OHLMANN T. (1990b), Affordances et vicariances mises en jeu par la régulation posturale. In Collectif (Eds.), *Informatique et différences individuelles.* Lyon, Presses Universitaires de Lyon.

OHLMANN T. (1991), La variabilité intra-individuelle provoquée : quelques pistes méthodologiques, *Actes des IXes journées de psychologie différentielle, 211-231.* Liège, Presses Universitaires de Liège.

OHLMANN T. (1995), Processus vicariants et théorie neutraliste de l'évolution : une nécessaire convergence. In J. Lautrey (Ed.), *Universel et différentiel en psychologie.* Paris, PUF.

OHLMANN T. (2000), *Contraintes situationnelles et plasticité individuelle*. Communication au colloque « Invariants et variabilité dans les sciences cognitives : compétences des systèmes vicariants ». Paris, 28 novembre.

PAOUR J.L., JAUME J. & DE ROBILLARD O. (1995), De l'évaluation dynamique à l'éducation cognitive, repères et questions. In F.P. Büchel (Ed.), *L'éducation cognitive, le développement de la capacité d'apprentissage et son évaluation*. (pp. 9-44). Lausanne, Delachaux et Niestlé.

PASQUIER D. (2003), *Test d'Evaluation Dynamique de l'Educabilité*, Paris, EAP.

PERLMUTTER M. & NYQUIST L. (1990), Relationship between self-reported physical and mental health and intelligence performance across adulthood. *Journal of Gerontology, 45,* 145-155.

PICHOT P. (1997, 15e éd. refondue), *Les tests mentaux*. Paris, PUF (Que sais-je ? n° 626).

PIÉRON H. (1963), *Examens et docimologie.* Paris, PUF.

POITRENAUD J, (1972). Structure des aptitudes cognitives et vieillissement. *Cahiers de la Fondation Nationale de Gérontologie,* 3-83.

RAVEN J. (1981), *Manuel des Matrices de Raven*. Issy les Moulineaux, EAP.

RAVEN J. (2001), Les matrices progressives de Raven : changement et stabilité à travers les cultures et le temps. In M. Huteau (Ed.), *Les figures de l'intelligence*. Paris, EAP.

REE M.J. & CARRETTA TR. (1998), General cognitive ability and occupational performance. ln, C.L. Cooper & L.T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology,* Volume 13. Wiley et Sons Ltd, Chichester, pp. 159-184.

REE M.J., EARLES J.A. & TEACHOUT M.S., (1994), Predicting job performance, not much more than g. *Journal of Applied Psychology*, 79, 518-524.

RÉMY L. & GILLES P-Y. (1999), Stratégies de résolution spatiale et numérique du D70. In M. Huteau & J. Lautrey (Eds), *Approches différentielles en Psychologie*. Rennes, P.U.R.

REMY L. (2001a), *Étude des stratégies de résolution d'une épreuve d'intelligence générale : variabilité intra-individuelle et différences interindividuelles.* Thèse de doctorat. Université de Provence Aix-Marseille I.

RÉMY L. (2001b), Les aptitudes des sujets sont-elles liées aux stratégies utilisées lors de la résolution d'une épreuve de facteur g ?, In A. Flieller, C. Bocéréan, J-L. Kop, E. Thiébaut, A-M. Toniolo et J. Tournois (Eds.), *Questions de psychologie différentielle.* Rennes, PUR.

RÉMY L. (2008), *Validation clinique de trois formes abrégées de la WAIS-III avec un échantillon de patients cérébro-lésés*. Communication aux XVIII[e] journées Internationales de Psychologie Différentielle, Université de Genève, 27-29 Août.

REUCHLIN M. (1978), Processus vicariants et différences individuelles, *Journal de Psychologie Normale et Pathologique, 2,* 133-145.

REUCHLIN M. (1991), *Les différences individuelles à l'école*. Paris, PUF.

REUCHLIN M. (1997), *La psychologie différentielle*. Paris, PUF.

REUCHLIN M. & BACHER F. (1989), *Les différences individuelles dans le développement cognitif de l'enfant*. Paris, PUF.

REY A. (1934), D'un procédé pour évaluer l'éducabilité. *Archives de Psychologie, XXIV* (96), 297-337.

RICHARD J-F. & AL. (1996), Le diagnostic cognitif, *Psychologie Française*, 41-1.

RICHARD J-F. (1996), Les différentes approches de l'analyse des compétences cognitives, *Psychologie Française*, 41, 1, 3-8.

RICHARD J-F., ZAMANI M. (1996), L'application des modèles de résolution de problèmes à l'analyse des tests, *Psychologie Française*, 41, 1, 77-88.

ROBERTSON I.T., SMITH J.M. (1989), Personal selection methods, in Robertson, I.T., Smith J.M. (Eds), *Advances in Selection and Assessment*, Wiley, Chichester.

ROBERTSON L.T., SMITH M. (2001), Personnel selection. *Journal of Occupational and Organizational Psychology,* 74, 441-472.

ROGERS W.A. & FISK A.D. (1999), Human factors, applied cognition, and aging. In F.I. Craik & T.A. Salthouse (Eds.), *The handbook of aging and cognition.* 2nd ed. Mahwah, NJ. Lawrence Erlbaum.

ROGOFF B. & WERTSCH J.V. (1984), *Children's learning in the "zone of proximal development"*. San Francisco, Jossey-Bass.

ROLLAND J-P. (2001), Comment évaluer un test ? In Levy Leboyer et al. (Eds.), R.H. *Les apports de la psychologie du travail*, p.35-52. Edition d'organisation.

ROLLAND J.-P. (2004), *L'évaluation de la personnalité. Le modèle des cinq facteurs.* Sprimont Belgique, Mardaga.

ROSENTHAL R & DIMATTEO M.R (2001), Meta-analysis, recent developments in quantitative methods for literature reviews. *Annual Review of Psychology,* 52, 59-82.

ROSENTHAL R. & RUBIN D.B. (1982), A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology,* 74, 166-169.

ROZENCWAJG P. (2003), Les stratégies de résolution de problèmes : une évaluation qualitative et intégrative de l'intelligence, *Psychologie et psychométrie, 24-4,* 37-62.

ROZENCWAJG P. (2005), *Pour une approche intégrative de l'intelligence, Un siècle après Binet.* Collection Mouvement des Savoirs. Paris, L'Harmattan.

ROZENCWAJG P. (2006), Quelques réflexions sur l'évaluation de l'intelligence générale : un retour à Binet ?, *Pratiques Psychologiques, 12, 3,* 395-410.

ROZENCWAJG P. & FRANCEQUIN G. (1999), Contributions de l'analyse des stratégies de résolution de problèmes à l'examen psychologique, *L'Orientation Scolaire et Professionnelle, 28, 1,* 63-82.

ROZENCWAJG P. & HUTEAU M. (1996), Les stratégies globale, analytique et synthétique dans les cubes de Kohs. *Psychologie Française*, 41, 1, 57-64.

ROZENCWAJG P., CHERFI M., FERRANDEZ A-M., LAUTREY J., LEMOINE C. & LOARER E. (2005), Age related changes in the strategies used by middle aged adults to solve a block design task. *The International Journal of Aging and Human Development, 60-2,* 159-182.

ROZENCWAJG P., CORROYER D. & ALTMAN P. (1999/2002), *Samuel : Diagnostic du fonctionnement cognitif (manuel),* Cergy, Delta Expert.

RUBTSOV V.V. (1981), The role of cooperation in the development of intelligence. *Soviet Psychology, 19* (4), 41-62.

S.F.P (2003), Recommandations Internationales sur l'utilisation des tests, *Pratiques Psychologiques*, numéro spécial hors série.

S.F.P (non daté), *La problématique de l'utilisation des tests*, document en ligne sur le site de la SFP www.sfpsy.org (consulté en mai 2008).

SALGADO J.F. (1999), Personnel selection methods. ln, C.L. Cooper, L.T. Robertson (Eds.), *International Review ofIndustrial and Organizational Psych*ology, Volume 14. Wiley et Sons Lld, Chichester, pp. 1-54.

SALGADO J.F. (2001), Pourquoi faut-il utiliser des épreuves d'Aptitude Mentale Générale en Recrutement ? In Levy-Leboyer et al. (Eds.), *R.H. Les apports de la psychologie du travail.* Edition d'organisation.

SALGADO LF., ANDERSON N., MOSCOSO S., BERTUA C., DE FRUYT F., ROLLAND J.P., 2003, A meta-analytic study of general mental ability validity for different occupations in the European Community. *Journal of Applied Psychology,* 88, 1068-1081.

SALOVEY P. & MAYER J. D. (1990), Emotional Intelligence. *Imagination, Cognition and Personality, 9* (3), 185-211.

SALTHOUSE T.A. (1994), The nature of the inflence of speed on adult age differences in cognition. *Psychological Review*, 30, 240-257.

SALTHOUSE T. A. (1996), The processing speed theory of adult age difference in cognition. Psychological Review, 103, 403-428.

SARRAZIN G. (Ed), (2003), *Normes de Pratique du Testing en psychologie et en éducation.* Montréal, Institut de recherches psychologiques.

SCHAIE K. W. (1979), The Primary Mental Abilities in adulthood, An exploration in the development of psychometric intelligence. In P.B. Baltes & O.G. Brim, Jr (Eds.), *Lifespan development and behavior* (Vol. 2). New York, Academic Press.

SCHAIE K. W. (1983), The Seattle Longitudinal Study, A 21-year exploration of psychometric intelligence in adulthood. In K.W. Schaie (Ed.), *Longitudinal studies of adult psychological development*. New York, Guilford.

SCHAIE K. W. (1994), The course of adult intellectual development. *American Psychologist, 49,* 304-313.

SCHAIE K.W. (1990), The optimization of cognitive functioning in old age, predictions based on cohort-sequential and longitudinal data. In P.B. Baltes & M.M. Baltes (Eds.), *Successful aging, perspectives from behavioral sciences* (pp. 94-117). New York, Cambridge University Press.

SCHAIE K.W. (1996), *Intellect development in adulthood*. The Seattle longitudinal study. Cambridge, Cambridge University Press.

SCHMIDT F.L., HUNTER J.E., (1998), The validity and utility of selection methods in personnel psychology, practical and theoretical implications of 85 years of research findings. *Psychological Bulletin,* 124, 262-274.

SCHMIDT F.L., HUNTER L, (1993), Tacit knowledge, practical intelligence, gencrai mental ability, and job knowledge. *Current Directions in Psychological Science* 2, 8-9.

SCHMIDT F.L., HUNTER J., PEARLMAN K., (1981), Task differences as moderators of aptitude test validity in selection, a red hening. *Journal of Applied Psychology,* 66, 166-185.

SCHORR D., BOWER G. H. & KIERNAN R. (1982), Stimulus variables in the block design task, *Journal of Consulting and Clinical Psychology, 50, 4,* 479-487.

SEWELL T.E. (1979), Intelligence and learning tasks as predictors of scholastic achievement in black and white first-grade children. *Journal of School Psychology, 17*, 325-332.

SEWELL T.E. (1987), Dynamic assessment as a nondiscriminatory procedure. In C.S. Lidz (Ed.) *Dynamic testing* (pp. 425-443). New York, Guilford Press.

SKA B., POISSANT A. & JOANETTE Y. (1997), La variabilité interindividuelle dans les modifications cognitives 1iées à *I'âge. L'Année Gérontologique, Numéro Spécial, Vieillir avec Succès.*

SNOW R.E., KYLLONEN P.C. & MARSHALEK B. (1984), The topography of ability and learning correlations. In R.J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (vol. 2, pp. 47-104). Hillsdale, NJ, Erlbaum.

SNOW R.E. & LOHMAN D.F. (1989), Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331).

SNYDERMAN M., ROTHMAN S. (1987), Survey of expert opinion on intelligence and aptitude testing. *American Psychologist*, 42, 2, 308-311.

SPEARMAN C.E. (1904), *General intelligence objectively measured and determined*. American Journal of Psychology, 15, 201-293.

SPEARMAN C.E. (1927), *The abilities of man, their nature and measurement*. New York, Macmillan.

SPEECE D.L., COOPER D.H. & KIBLER J.M. (1990), Dynamic testing. Individual differences and academic achievement. *Learning and Individual Differences, 2,* 113-127.

SPELBERG H. (1987), Problem-solving strategies on the block-design task, *Perceptual and Motor Skills, 65,* 99-104.

STERNBERG R. J. (1985), *Beyond IQ, A triarchic theory of human intelligence.* New York, Cambridge University Press.

STERNBERG R. & DETTERMAN D. (1986), *What is intelligence*. New Jersey, Ablex Publishing Corporation.

STERNBERG R.J., FORSYTHE G.B., HEDLUND J.H., HORVATH J.A., WAGNER R.K., WILLIAMS W.M., SNOOK S.A. & GRIGORENKO E.L. (2000), *Practical Intelligence in Everyday Life*. New York, Cambridge University Press.

STERNBERG R.J., GRIGORENKO E.L. & JARVIN L. (1997), School-based tests of the triarchic theory of intelligence, three settings, three samples, three syllabi. *Joumal of Educational Psychology.*

TERMAN L.M., (1916), *The measurement of intelligence*. Boston, Houghton Mifflin.

THIEBAUT E. (2000), *L'intelligence et sa mesure. Introduction aux tests de Bonnardel*. Paris, EAP.

THIEBAUT E. & BIDAN-FORTIER C. (2003), Manuel de la batterie NV5-R. Paris, EAP.

THIÉBAUT E. & RICHOUX V. (2005), Eléments de validité prédictive des scores à la batterie d'aptitudes cognitives NV5-R, *Pratiques Psychologiques, 11, 404-416.*

THORNDIKE E. L. (1920), *Intelligence and its use*. *Harper's Magazine, 140*, 227-235.

THURSTONE L.L. (1935), *The vectors of the mind* Chicago, Univ. Chicago Press.

THURSTONE L.L. (1938), *Primary Mental Abilities*. Chicago, Chicago University Press.

THURSTONE L.L. & THURSTONE T.G. (1941), Factorial studies of intelligence, *Psychometric Monographs*, n°2.

TORT M. (1974), *Le Quotient Intellectuel*. La Découverte, Paris, Maspero.

ULLMO J. (1969), *La pensée scientifique moderne*. Paris, Flammarion.

VAN DE VIJVER F. & POORTINGA Y. (1997), Towards an Integrated Analysis of Bias in Cross-Cultural Assessment. *European Journal of Psychological Assessment, 13,* 29-37.

VERNON P.E. (1950), *Structure of human abilities*. London, Methuen.

VERNON P.E. (1952), *La structure des aptitudes humaines*, Paris, PUF.

VIGNEAU F., DOUGLAS A. B. & STOKES T. L. (2001), La multidimensionnalité d'un test de facteur g ? Vers une approche expérimentale du test des Matrices de Raven, In A. Flieller, C. Bocéréan, J-L. Kop, E. Thiébaut, A-M. Toniolo et J. Tournois (Eds.), *Questions de psychologie différentielle.* Rennes, PUR.

VOM HOFE A. & LEVY-LEBOYER C. (1993), Evaluation of the use of personality tests in personel selection in france. *Revue Européenne de Psychologie Appliquée, 43* (3), 221-227.

VRIGNAUD P. (1994), Méthodologie de l'évaluation. In M. Huteau (Ed.) *Actes du Colloque international "Les techniques psychologiques d'évaluation des personnes"* (pp. 62-67). Issy-les-Moulineaux, EAP.

VRIGNAUD P. (1996), Les tests au XXI^{e} siècle. Que peut-on attendre des évolutions méthodologiques et technologiques dans le domaine de l'évaluation psychologique des personnes ? *Pratiques Psychologiques, 4*, 5-27.

VRIGNAUD P. (1996), Les tests au XXIème siècle. Que peut-on attendre des évolutions méthodologiques et technologiques dans le domaine de l'évaluation psychologique des personnes ? Pratiques Psychologiques, 4, 5-27.

VRIGNAUD P. (2000), Psychological Assessment, An Overview of French-Language Theory and Methods. In M. R. Rozenzweig & K. Pawlik (Eds). *The International Handbook of Psychology*. (pp. 387-392). London, Sage.

VRIGNAUD P. (2001), Évaluation sans frontières : comparaisons interculturelles et évaluations dans le domaine de la cognition, In M. Huteau, Les figures de l'intelligence. Paris, EAP.

VRIGNAUD P. (2002a), Les biais de mesure, savoir les identifier pour y remédier. *Bulletin de Psychologie*, *55(6)*, 625-634.

VRIGNAUD P. (2002b), Psychométrie et validation de la mesure. In A. Vallet, G.Bonnet, J.-C. Emin, J. Levasseur, T. Rocher, A. Blum, F. Guérin-Pace, P. Vrignaud, X. d'Haultfoeuille, F. Murat,D. Verger, P. Zamora (Eds), *Enquête méthodologique « Information et Vie Quotidienne ». Tome 1 : Bilan du test – 1*. Collection Méthodologie Statistique de l'INSEE, 0202. pp 35-49. Paris, Institut National de la Statistique et des Etudes Economiques.

VRIGNAUD P. & BERNAUD J-L., Eds (2005), *L'évaluation des intérêts professionnels.* Paris, Mardaga.

VRIGNAUD P. & CHARTIER P. (2003), Apport de l'analyse des séquences à l'étude des processus cognitifs, *Psychologie et Psychométrie, 24(4),* 77-114.

VRIGNAUD P. & LOARER, E. (2008), Tests et recrutement. In S. Ionescu et A. Blanchet (Eds.) *Nouveau cours de psychologie, volume « Psychologie sociale et ressources humaines »* coordonné par M. Bromberg et A. Trognon. Paris, PUF.

VRIGNAUD P. (2003), Objectivité et authenticité dans l'évaluation. Avantages et inconvénients des Questions à Choix Multiples et des Questions à Réponses Complexes pour l'évaluation des compétences verbales. *Psychologie et Psychométrie, 24, 2/3,* 147-188.

VYGOTSKI L.S. (1934/1985), *Pensée et langage*. Paris, Editions Sociales.

VYGOTSKI L.S. (1985), Le problème de l'enseignement et du développement mental à l'âge scolaire. In B. Schneuwly & J.P. Bronckart (Eds.), *Vygotski aujourd'hui.* Neuchatel - Paris, Delachaux et Niestlé.

WAGNER R.K. & STERNBERG R.J. (1984), *Alternative conceptions of intelligence and their implications for education, Review of educational Research*, 54(2), 179-223.

WECHSLER D. (1939), *The measurement of adult intelligence*. Baltimore, Williams & Wilkins.

WECHSLER D. (1956), *La mesure de l'intelligence de l'adulte.* Paris, PUF.

WESCHLER D. (1958), *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore, MD, The Williams & Wilkins Company.

WECHSLER D. (1996), *Manuel du WISC-III*. Paris, ECPA.

WECHSLER D. (2000), *Manuel de la WAIS-III*. Paris, ECPA.

WECHSLER D. (2005a), *WISC-IV. Manuel d'administration et de cotation*. Paris, ECPA.

WECHSLER D. (2005b), *WISC-IV. Manuel d'interprétation*. Paris, ECPA.

WERTSCH J.V. (1984), The zone of proximal development, some conceptual issues. In B. Rogoff & J.V. Wertsch, (Eds.), *Chidren's learning in the zone of proximal development* (pp. 7-18). San Francisco, Jossey-Bass.

WERTSCH J.V. & TULVISTE P. (1992), L.S. Vygotsky and contemporary developmental psychology. *Developmental Psychology, 28*, 548-557.

ZAZZO R., GILLY M. & VERBARAD M. (1966), *Nouvelle échelle métrique de l'intelligence*. Paris, Colin.

ZURFLUH J. (1976), *Les tests mentaux*. Paris, Delarge.

LES OUTILS DU PSYCHOLOGUE

Philippe Chartier
Even Loarer

ÉVALUER L'INTELLIGENCE LOGIQUE

Choix des épreuves • Passation
Interprétation • Restitution

Conçu comme un support à la pratique, cet ouvrage rappelle les règles et les précautions à prendre pour parvenir à une mesure valide de l'intelligence. Il propose une synthèse des différentes approches de l'intelligence logique et décrit, avec de nombreux exemples, les méthodes utilisées pour la mesurer.

Il présente également des approches plus récentes de l'évaluation : l'analyse des stratégies de résolution et la mesure d'un potentiel d'apprentissage.

Chacun des grands types d'épreuves fait l'objet :

- d'une description de son cadre théorique et des conditions de sa validité ;
- d'une analyse détaillée de sa construction ;
- d'indications pour sa mise en œuvre, son interprétation et sa restitution.

Cet ouvrage est principalement destiné aux psychologues praticiens, ainsi qu'aux étudiants en psychologie soucieux de se former à la pratique de l'évaluation des capacités intellectuelles dans le respect des règles de déontologie et de validité scientifique.

Échelles d'intelligence. Les échelles de Wechsler :
- **WISC-III, WISC-IV, WAIS-III.**

Tests de facteur g
- **Matrices de Raven, NNAT, D48, D70 et D2000, R85/R2000.**

Batteries factorielles
- **NV7, NV5-R, DAT 5.**

Épreuves de Potentiel d'apprentissage :
- **le TEDE 6.**

Analyse des stratégies :
- **le logiciel SAMUEL.**

PHILIPPE CHARTIER
est maître de conférences en psychologie différentielle à l'INETOP (CNAM, Paris).

EVEN LOARER
est professeur de psychologie du travail à l'Université Paris Ouest-Nanterre La Défense.

ISBN 978-2-10-053535-4

www.dunod.com

DUNOD