Docimologie critique: des difficultés de noter des copies et d'attribuer des notes aux élèves

Dieudonné Leclercq, Julien Nicaise, Marc Demeuse

DOCIMOLOGIE CRITIQUE : DES DIFFICULTÉS DE NOTER DES COPIES ET D'ATTRIBUER DES NOTES AUX ÉLÈVES[1]

Dieudonné Leclercq
Julien Nicaise
Marc Demeuse

1. L'histoire de la problématique

L'école est un lieu dans lequel l'évaluation est omniprésente. Il semble même parfois à certains que l'élève fréquente davantage l'école pour récolter des notes que pour apprendre véritablement quelque chose.

Avec la massification et la démocratisation de l'accès à tous les niveaux scolaires, il faut pouvoir comptabiliser échecs et succès à travers un système de notation de façon à rendre un "verdict" en fin d'année. La notation est ainsi une réponse à la division du travail et à l'unicité de l'élève et du temps: il doit être possible de prendre, à un moment donné, une décision relative à chaque élève, ce qui implique la prise en compte d'informations provenant de sources multiples.

Si donc l'école a vu se systématiser et se professionnaliser "l'art de bien enseigner" à travers la didactique, elle a connu le même développement de "l'art de bien évaluer" à travers la docimologie. Dans la suite du texte, nous envisagerons principalement les travers de l'évaluation classique des élèves, nous en pointerons les limites, en suivant les chemins de la docimologie critique. Nous ne développerons donc pas les solutions et les remèdes. D'autres cours existent dans le cursus des étudiants - principalement de ceux qui suivront la formation de la licence en sciences de l'éducation - qui leur permettent d'envisager les améliorations possibles. Ce chapitre ne constitue donc qu'une introduction critique à la docimologie pratique, la problématique étant d'être conscient et attentif aux difficultés rencontrées lors de l'évaluation.

Les développements de la docimologie et de la mise en cause des notes scolaires remontent au début du vingtième siècle. Dès les années 1910, les Etats-Unis firent confiance aux QCM dans les tests de sélection, par souci d'objectivité et en réponse à la difficulté de noter. A partir de 1922, en France, Henri Piéron s'attaque aux problèmes posés par la subjectivité de la note. Dès 1929, il attire l'attention en ces termes : *« C'est un principe général que, pour être reçu à un examen, il faut avoir la moyenne, ...dès lors, ...pour un grand nombre de candidats,*

[1] Ce chapitre résulte d'une synthèse de différents documents:
Leclercq, D. (1999). Chapitre 3 - Les production de "synthèse" et la docimologie critique. In *Edumétrie et Docimologie.* Université de Liège.
Nicaise, J. (2001). *Pratiques, sens et sens pratique au cœur des évolutions institutionnelles: les instituteurs de sixième primaire et le jugement professoral.* Université de Liège: mémoire de licence (non publié).
Nicaise, J. (2002). De la non-objectivité du jugement professoral en matière d'évaluation des performances des élèves. *Les Cahiers du Service de Pédagogie expérimentale, 11-12.*
Il s'inspire aussi très largement de l'ouvrage de G. de Landsheere (1971) qui devient malheureusement très difficile de se procurer.

ce sera …le hasard qui décidera de leur admission ou de leur recalage. En effet, on sait que… c'est dans la région moyenne qu'ils se massent… ». (Piéron, 1963, p.9)

Aux USA, en Angleterre, et même en Belgique, diverses expériences mettent en évidence le manque de fiabilité des notes scolaires. Piéron (1963, p. 13) cite notamment les travaux menés à ce sujet, en 1931, par Andréa Jadoulle, la célèbre psychopédagogue du *Laboratoire de pédagogie d'Angleur*[2]. En France, Laugier et Weinberg étudient ce même phénomène dès 1927.

C'est en 1931 qu'une impulsion déterminante sera donnée par la subsidiation, par la *Carnegie Corporation* de New York, d'une recherche pilotée par *l'International Institute of Education* du *Teacher's College* de l'université Columbia, et fonctionnant via des commissions nationales : américaine, anglaise, écossaise, finlandaise, française, suisse et allemande (cette dernière étant arrêtée par la nazification de 1933).

La commission française utilisa des copies notées lors du fameux baccalauréat. En 1934 fut publié, par la commission française, le rapport « Etudes docimologiques » (Laugier, Piéron, Pieron, Toulouse et Weinberg, 1934). Le baccalauréat offrait une situation exceptionnelle puisque les mêmes questions sont posées à de très nombreux étudiants, durant de véritables examens, et sont collectées et corrigées par de nombreux correcteurs sélectionnés. A travers cette situation, complètement externe (encart 1), l'équipe française met en évidence de nombreux biais de notation. On se doute que la guerre interrompit ce processus de recherche sur le continent européen.

Encart 1 - *Epreuves internes et épreuves externes*

a) Les examens internes

En Belgique, quel que soit le niveau d'enseignement considéré, les examens sont généralement administrés par les enseignants qui ont donné les cours. C'est eux qui créent les questions et qui corrigent les copies. Cette façon de faire a des avantages comme celui de « coller » à la matière qui a effectivement été enseignée, ou celui d'une familiarité des élèves au type de questions. Il a le désavantage de laisser libre cours aux différences (de sévérité par exemple) intercorrecteurs ou interétablissements, ce qui pose le problème éthique de l'équité et de l'égalité de traitement, tout spécialement quand l'examen est « sanctionnant » et quand le professeur sait de qui il corrige la copie.

b) Les examens externes à correction subjective

Dans un souci d'égalité de traitement, la France, depuis Napoléon, pratique le baccalauréat, examen (le même pour tous les étudiants d'une même « Académie ») conçu et corrigé par des enseignants n'ayant pas participé à l'encadrement des candidats évalués dont les copies sont rendues anonymes. On devine les précautions à prendre par les formateurs pour respecter « le programme » et par les concepteurs des questions pour « éviter les fuites ». Ces examens restent toutefois subjectifs quant à la correction.

c) Les examens externes à correction objective

Poussant encore plus loin le souci d' « *equity* » et de « *unbiased evaluation* », les Américains ont non seulement conçu, à la charnière du secondaire et de l'enseignement supérieur, des examens (par exemple le *Scholastic Aptitude Test* ou *SAT*) qui sont les mêmes pour tous, mais dont la correction est objective (ce sont des QCM). D'où l'expression *« objective tests* », par un élargissement de sens légèrement abusif.

Le terme "docimologie" apparaît quant à lui en 1929 sous la plume d'Henry Piéron et est popularisé par celui-ci dans son ouvrage intitulé "Examens et docimologie", en 1963. Ce mot puise sa racine dans le grec (examiner, épreuve). A ses débuts, la docimologie est surtout

[2] *« En Belgique où, sur l'initiative d'un échevin éclairé, René Jadot, avait été fondé à Angleur un laboratoire de psychopédagogie, des expériences avaient été faites par Mlle Jadoulle en 1931, confiant à 4 correcteurs le soin de noter des compositions (problèmes et questions relatives à l'intelligence d'un test) en 2^{e} et en 6^{e} année. Les conclusions étaient très pessimistes, un élève se trouvant classé 6^{e}, 14^{e} ou 23^{e} sur une trentaine. »* (Péron, 1963, p. 13).
En Belgique, on doit aussi à G. de Landsheere un ouvrage intitulé *Evaluation continue et examens: précis de docimologie,* publié à Bruxelles, chez Labor, et à Paris, chez Nathan, en 1971. Cet ouvrage a été publié et republié pendant plus de 20 ans.

critique ou négative: elle met en évidence les problèmes, sans les résoudre, du moins de manière pratique, au niveau où le problème se pose, c'est-à-dire au niveau des enseignants chargés de procéder à l'évaluation. Progressivement cependant, les chercheurs s'attachent à proposer des solutions qui permettent de limiter au mieux le caractère subjectif de la notation. Comme nous allons le voir par la suite, ce caractère subjectif n'est pas à imputer uniquement au maître chargé de noter l'élève, il relève de mécanismes souvent complexes et qui mettent en jeu enseignants, élèves et système éducatif.

Les méthodes employées pour étudier les biais de notation reposent sur différentes procédures, selon le type de biais à mettre en évidence. On peut ainsi utiliser, de manière expérimentale, les approches suivantes:

(a) Une même série de copies est corrigée plusieurs fois par le même correcteur, à des moments différents, sans que celui-ci s'en rendre compte, ce qui permet de mesurer la ***stabilité intra-correcteurs***;

(b) Une même série de copies est corrigée par plusieurs correcteurs différents, ce qui permet de mesurer la ***concordance inter-correcteurs***;

(c) Une même copie est placée dans un ensemble de copies dans des positions différentes (précédée de copies meilleures ou plus faibles), ce qui permet de mesurer l'effet de contraste, ou de séquence;

(d) Une même copie est placée dans un ensemble de copies dont les valeurs sont plus ou moins dispersées largement (tantôt parmi des copies ayant toutes reçu la même note lors d'une évaluation préalable, tantôt parmi des copies très variées en qualité); etc.

(e) Une même copie est corrigée par plusieurs groupes de correcteurs auxquels on fournit des informations complémentaires différentes sur l'élèves, ses notes antérieures…

2. Les trois sources « d'erreur »

Si l'on cherche à identifier les sources d'erreurs ou de biais qui entachent la notation par les enseignants, pour conserver une approche comparable à celle adoptée par la théorie classique des tests, les chercheurs identifient au moins trois sources principales : les enseignants, le système scolaire et les élèves. Nous allons aborder successivement ces trois sources. Nous montrerons combien il importe d'envisager un problème qui semble assez singulier, donner une note à un élève particulier, à travers un ensemble beaucoup plus large de déterminants et d'interactions (Perrenoud, 1989, 1998).

2.1. Le système scolaire

La première source de biais que l'on doit prendre en considération est l'influence du contexte scolaire sur les procédures d'évaluation en général. Ainsi, **la classe** dans laquelle se trouve l'élève peut être déterminante. Les conclusions tirées notamment par Grisay sur « l'effet Posthumus » au début des années quatre-vingt – et de nombreuses fois vérifiées et commentées par la suite (Grisay, 1984 ; Crahay, 1996 ; Demeuse, 2002) – offrent un très large aperçu sur les interférences que peut avoir le contexte d'une « classe particulière » sur l'évaluation des élèves qui la composent : avec les mêmes performances, et toute autre chose restant égale par ailleurs, un élève est jugé par son professeur comme un « bon élève » dans une classe alors qu'il peut se voir contraint de doubler son année scolaire dans une autre ! Tout dépend non pas des performances particulières de l'élève dans l'absolu mais bien de ses performances par rapport à celles de ses condisciples. Le hasard du microcosme de la classe dans laquelle se trouve les élèves est donc primordiale puisque le professeur est fréquemment

poussé à établir des différences de performance entre eux, et ce même si ceux-ci sont très proches, c'est-à-dire à adopter une attitude plus normative que critériée. De ce fait, leurs résultats en fin d'année sont souvent distribués selon une courbe de Gauss (quelques élèves « faibles », quelques élèves « forts », et la grande majorité dans la « moyenne ») (Perrenoud, 1995 [1984] ; Crahay, 1996 ; Merle, 1998).

Cependant si la classe dans laquelle est placé un élève plutôt que dans une autre et son influence sur l'évaluation finale est importante, **l'établissement scolaire** l'est parfois tout autant. Alors que de nombreux travaux – notamment américains– et leurs diverses interprétations avaient pu laisser sous-entendre dans la foulée des *Golden sixties* (Coleman et al., 1966 ; Jencks, 1979 [1972]), que l'école a peu d'impact sur les résultats des élèves, le « *school can make a difference* » est actuellement reconnu, aussi bien au niveau des résultats effectifs des élèves que de l'évaluation de ceux-ci.

Il semble toute fois que l'*effet-classe* que nous venons d'évoquer soit bien plus important que l'*effet-établissement* (Bressoux, 1994 ; *id.*, 1995), même si l'influence de ce dernier est pourtant indéniable. Cette différence entre établissements peut résulter d'une stratégie délibérée destinée à pratiquer une sélection par écrémage et/ou à médiatiser une certaine image de l'excellence dont la « fameuse réputation » de certaines écoles d'élite est la concrétisation (Duru-Bellat et Mingat, 1993 ; Merle, 1998).

Et quant bien même la complexité du contexte de l'école et de la classe ne suffirait pas à laisser entrevoir toute l'incertitude, la non-objectivité dont est déjà empli à ce stade l'acte évaluatif de l'enseignant, des biais bien plus élémentaires peuvent démontrer que la note est encore parfois influencée par d'autres déterminants totalement extérieurs au travail de l'élève en soi : depuis longtemps par exemple la docimologie a pu montrer qu'une même copie est notée différemment par l'enseignant selon son **ordre de correction**, selon qu'elle se trouve parmi les premières ou parmi les dernières feuilles de la pile que l'enseignant doit corriger (Bonniol, 1965) ; et dans le même ordre d'idée, qu'une même copie sera jugée différemment selon qu'elle suit une copie jugée « très bonne » ou « très faible » par l'enseignant (Bonniol et Piolat, 1971). Bon nombre d'enseignants n'hésitent d'ailleurs pas à déclarer qu'ils établissent leur barème de notation seulement après avoir lu plusieurs copies, et que celui-ci peut être appliqué différemment selon les élèves et selon la succession des résultats, copie après copie (Chevallard, 1991 ; Nicaise, 2001).

2.2. Les élèves

A côté de ce contexte de scolarisation, des particularités intrinsèques à l'élève peuvent également influencer subjectivement le jugement professoral : dans la foulée des premières sociologies dénonciatrices des années soixante et septante (Bourdieu et Passeron, 1964 ; *id.*, 1970 ; Baudelot et Establet, 1971 ; Boudon, 1973), on pensera avant toute autre chose à **l'origine sociale** des élèves et aux stéréotypies diverses qui peuvent y être associées : ainsi, certains correcteurs ont parfois tendance à attribuer de meilleures notes aux enfants issus des milieux les plus favorisés (Pourtois et al., 1978) alors que dans d'autres circonstances, ce sont justement les élèves issus des milieux défavorisés qui sont « surcotés », notamment pour des raisons de « paternalisme bienveillant » et de correction – si minime soit-elle – volontaire des inégalités sociales et scolaires (Dardenne, 1999 ; Nicaise, 2001). Le même type de conclusions a également pu être tiré dans ce sens avec des déterminants comme **l'apparence physique** ou **le genre sexuel** des élèves : parfois les élèves jugés « plus beaux » ou « plus proches des idéaux médiatiques » par les enseignants peuvent être mieux cotés (Leyens et Yzerbit, 1997 ; Merle, 1998) et, dans le même ordre d'idée, il semble que les filles ont à souffrir plus fréquemment que les garçons quant à leur notation, notamment suite aux

phénomènes de « menace stéréotypique » et de reproduction des inégalités sociales en matière de genre sexuel (Dardenne, 1999). Par contre, il peut apparaître également dans d'autres circonstances que les mêmes filles sont habituellement mieux évaluées parce qu'elles font preuve d'un « meilleur comportement » en classe et maîtrisent mieux les règles et exigences du « métier d'élève » – soit donc qu'elles sont plus proches que leurs pendants masculins d'un comportement idéalement attendu par le professeur (Felouzis, 1993 ; Duru-Bellat, 1995). Par là, on peut constater que l'évaluation des enseignants porte tout autant sur les « savoir-être » que sur les plus classiques savoirs et savoir-faire (Bourdieu et Passeron, 1970 ; Crahay, 2000), et que le même incitant peut influencer celle-ci dans des sens totalement opposés.

Comme le complexe « scolarité », le complexe « élève » est très influent et la relation ne cesse de se complexifier encore un peu plus lorsque entrent en jeu des caractéristiques qui sont simultanément dépendantes des deux : pensons par exemple au **niveau scolaire** de l'élève (les résultats de ses évaluations précédentes, notamment ceux présentés dans son bulletin scolaire) et au **statut de sa classe** (la réputation, l'image véhiculée par la classe qu'il fréquente, notamment celle présentée lors des conseils de classe). Par un comportement proche d'un classique phénomène de réduction de la dissonance cognitive (Festinger, 1957), un enseignant peut parfois être tenté de minimiser la différence qui apparaîtrait entre sa propre attente quant aux résultats de l'élève et les résultats effectivement obtenus. L'habituel « premier de classe » peut se voir ainsi tiré vers le haut alors que les performances réalisées ne correspondent pas à ce que l'enseignant attendait de lui comme à l'accoutumée (Caverni, Fabre et Noizet, 1975, Merle, 1998).

Le complexe « élève » peut être vu comme une mise en commun d'une multitude d'effets d'attente qui se rapportent tous, de près ou de loin, à la thèse devenue classique aujourd'hui de « Pygmalion à l'école » de Rosenthal et Jacobson . Selon ce modèle, certains déterminants (origine sociale, sexe, niveau scolaire, etc.) amènent l'enseignant à développer rapidement des attentes diverses vis-à-vis de ses élèves (notamment sur leurs résultats scolaires futurs) et à différentier peu à peu, son propre comportement - verbal comme non-verbal, conscient comme inconscient. Il a tendance à orienter ses élèves vers le résultat scolaire attendu : qu'elle soit positive ou négative, la prédiction peut alors se révéler fortement créatrice d'effets (Rosenthal et Jacobson, 1969 ; Good, 1987).

2.3. Les enseignants

Comme nous l'avons souligné précédemment, l'existence et l'interférence de nombreux biais dans la procédure évaluative peut encore être exacerbée puisque, dans ce domaine tout du moins, la liberté du maître au sein de sa classe est presque totale : il administre les épreuves comme il l'entend, il choisit la matière et le sujet des interrogations, il choisit leur forme, leur moment, leur durée, leur importance sur la note finale, il choisit les critères et les normes qui détermineront son jugement professoral, ... puis c'est lui qui applique le modèle d'évaluation aux productions de ses élèves qui s'avèrent, d'une certaine manière, le reflet de son propre travail d'enseignant.

La quête de l'objectivité de la note est donc semée d'embûches majeures. Dépassant par là la vision mécaniste du maître simple notateur dans l'absolu, il faut appréhender que l'acte évaluatif passe également au travers de nombreux « filtres interactifs » avant que la note finale puisse être arrêtée. Donc, outre les biais classiques que nous avons présentés, le fait de noter un élève est également une action proprement rationnelle qui trouve ses fondements – pour reprendre le raisonnement de Weber (1971) – à la fois dans les intérêts et les valeurs propres de l'enseignant. Ainsi, si le jugement professoral peut être dépendant de l'origine sociale des élèves, de leur âge, de leur sexe ou du type d'établissement fréquenté, il l'est tout

autant – comme le souligne très judicieusement Merle (Merle, 1996) – d'un ensemble quotidien « d'arrangements » et de « bricolage » des notes, intentionnels ou non[3]. Dès lors, la compréhension des actes évaluatifs nécessite d'une vision ultra-systémique (Perrenoud, 1995, 1998), mais se double de l'exigence d'une approche ultra-individuelle et biographique de chaque sujet évaluateur particulier.

2.3.1. Les arrangements internes

Le premier type de « bricolages », de modifications plus ou moins licite des procédures d'évaluation, est destiné directement à la classe *in vivo* et aux élèves qui la composent. Il peut servir à entretenir un bon climat de travail, à encourager les élèves qui éprouvent des difficultés ou qui ont des problèmes d'ordre extrascolaire (dans ce cas, les notes sont revues « à la hausse »), à restaurer l'autorité concrète ou symbolique du maître en sanctionnant certains comportements (les notes sont alors revues « à la baisse »), à sauvegarder une moyenne de points habituelle, à amener un élève vers une orientation future plutôt qu'une autre, à céder aux éventuelles « pressions » diverses des élèves, etc. Habituellement, ces comportements ne « sortent » pas de la classe, ils ne sont pas délibérément cachés par le maître mais celui-ci s'en vante rarement car ils font partie de sa propre « cuisine interne », de ses procédures personnelles (Merle, 1996). L'apposition d'une note relève donc bien également de la transaction, et constitue un moment particulier – mais essentiel – d'un processus beaucoup plus large, celui d'une véritable « négociation didactique » entre l'enseignant et ses élèves (Chevallard, 1991).

2.3.2. Les arrangements externes

Les arrangements dits externes prennent la même forme que les précédents, mais ils sont destinés à la direction de l'école, à l'administration, aux collègues, aux parents d'élèves, bref, à toute personne qui ne participe pas directement au quotidien de la classe, mais qui interagit néanmoins avec elle. Il s'agit souvent pour l'enseignant de présenter une image de sa classe qui satisfasse au mieux ces personnes extérieures : qu'adviendrait-il si trop d'élèves étaient en échec ? Qu'adviendrait-il si tous avaient des résultats exceptionnels ? La réputation et le « statut » prêté à l'enseignant pourrait être mis à mal et il en serait de même pour celui de l'établissement.

Ce type d'arrangements est évidemment lié très étroitement au précédent (Merle, 1996) : si par exemple une interrogation écrite est particulièrement mal exécutée par l'ensemble de la classe et que le professeur décide de ne pas en tenir compte, la finalité d'un tel acte est double car celui-ci a des conséquences internes, mais aussi externes à la classe considérée comme le seul groupe d'élèves.

Au travers de ces interactions, une évaluation trop sévère apparaît très vite comme injuste, mais une évaluation trop généralement favorable s'apparente à du laxisme et celui-ci nuit inévitablement à la réputation de l'enseignant et de l'établissement (Perrenoud, 1998). Un difficile équilibre doit donc s'établir. Ainsi, l'enseignant tente souvent de ne pas s'écarter de ses moyennes et des distributions habituelles des notes (Grisay, 1984 ; Crahay, 1996). Il montre ainsi à quiconque qu'il « tient » sa classe (Chevallard, 1991). Sur l'importance de ces arrangements externes – soit pour l'essentiel « intéressés » –, Grisay montre que de nombreux enseignants, une minorité il est vrai, avouent prendre en considération des éléments « *illégitimes* » très diversifiés pour établir leur décision finale – de réussite ou de doublement – pour certains de leurs élèves. Ainsi, les instituteurs peuvent être influencés par les

[3] On admettrait assez difficilement que tous les élèves obtiennent le maximum et encore moins que tous échouent lors d'une évaluation, d'un examen. Il existe donc assez généralement des procédures, conscientes ou non, d'ajustement de la distribution des résultats bruts à un modèle acceptable.

insistances répétées des parents, par le risque que ceux-ci retirent leur enfant de l'établissement si celui-ci venait à doubler, par le fait que la réussite – ou le doublement – d'un élève provoque l'ouverture ou la fermeture d'une classe à la future rentrée scolaire, par le fait que le prochain enseignant de l'élève sera prêt à l'aider et le soutenir plus qu'à l'accoutumée, par le fait également que l'école accorde ou non une grande importance à sa « réputation », etc. (Grisay, 1991 [1986]). Il est évident qu'ici encore les enseignants qui usent de ces pratiques ont de grandes réticences à les dévoiler expressément : « *faire part de sa "cuisine" évaluative et partager ses doutes nécessitent de dévoiler les limites de son propre jugement et de se mettre en cause professionnellement* » (Merle, 1996 : 86).

2.3.3. Les arrangements pour soi[4]

Ce troisième type « d'arrangements évaluatifs » est fréquemment ignoré dans de nombreuses études et ceci principalement à cause de la difficulté de les appréhender et de les regrouper au sein de types-idéaux exploitables puisqu'ils dépendent directement de l'histoire et de la personnalité même du sujet-correcteur. Ils sont pris à l'égard de soi-même et peuvent dépendre d'une foule de représentations personnelles, chacune plus difficilement saisissable que l'autre : « l'idéal pédagogique » de l'enseignant, sa conception général de l'éducation, son propre parcours scolaire, son origine sociale, ses engagements politiques et associatifs particuliers, etc. Les normes de justice scolaire auxquelles peut adhérer, parfois avec force, l'enseignant sont également déterminantes : l'égalité des chances, l'égalité de traitement ou l'égalité de résultat, le besoin de l'élève, sa contribution et son mérite, le refus de doublement, etc. (Barrère, 2000 ; Nicaise, 2001). Il en est de même avec les conceptions générales sur les écoliers : sont-ce d'abord des élèves à scolariser ou des enfants à éduquer et à socialiser que l'on a en face de soi ? – (Dubet et Martuccelli, 1996). Tous ces facteurs, tous ces sens donnés aux pratiques par le sujet selon ses représentations interviennent évidemment de façon conjointe.

2.4. En bref…

Ces différentes pratiques montrent donc bien que l'évaluation des élèves, avant d'être une simple apposition d'une note que l'on croit encore parfois être « vraie », « *relève de processus et de procédures au croisement des contraintes sociales et des biographies des élèves et des maîtres* » (Merle, 1996, p. 306). Ainsi, le processus d'évaluation est dépendant d'un triple rapport entre le professeur et ses élèves, le professeur et ses contraintes externes, et le professeur et son passé, son intériorité, lorsqu'il s'engage personnellement dans son travail.

La procédure évaluative n'est donc pas un acte identiquement posé par chaque enseignant. L'objectivité n'y est pas une norme. Ses règles et ses critères, même les plus généraux, ne sont d'ailleurs que trop rarement définis et arrêtés au préalable, comme ce peut être le cas d'une procédure juridique par exemple.

[4] Le terme « d'arrangements » utilisé ici dépasse donc de très loin le sens littéral des notions de « négociations » (Chevallard, 1991) ou de « stratégies » (Perrenoud, 1995 [1984]), également utilisées dans le contexte des interactions scolaires puisque ces « arrangements » sont également réalisés « pour soi », « *parce que tel comportement évaluatif s'inscrit dans son histoire personnelle et s'impose comme une exigence* » (Merle, 1996, p. 76).

3. Quelques exemples de biais mis en évidence par la docimologie critique

3.1. La distribution forcée

On attribue assez généralement à ***Posthumus***, enseignant hollandais en poste en Indonésie durant la seconde guerre mondiale et interné dans un camp japonais durant celle-ci, la paternité d'une loi formulée de la manière suivante par De Landsheere (1992, p. 242):
« Un enseignant tend à ajuster le niveau de son enseignement et ses appréciations des performances des élèves de façon à conserver, d'année en année, approximativement la même distribution (gaussienne) de notes[5]*. »*

Cette "loi de Posthumus" indique que la distribution des notes résulte d'une sorte de prototype, communément admis: il existe peu d'élèves exceptionnels (très faibles ou très brillants), mais beaucoup d'élèves relativement moyens. Lorsque l'on se place dans la situation d'une épreuve interne, c'est assez souvent ce type de distribution qui est mise en évidence, alors même que des épreuves externes, appliquées aux mêmes élèves, indiquent des niveaux moyens très variables et des dispersions différentes, d'une classe à l'autre.

De Landsheere (1992, p. 36), explique ce phénomène de la manière suivante :

« Un professeur qui enseigne de façon non individualisée dans une classe où les élèves ne sont pas spécialement sélectionnés donne normalement à son cours un degré de difficulté adapté à la majorité du groupe. Si l'ajustement est correct, il y aura donc beaucoup de résultats moyens, peu de très bons et peu de très mauvais. La distribution de ces résultats s'approchera de la courbe gaussienne. Cette distribution, dite normale, est à l'image de beaucoup de qualités humaines, telles qu'elles se répartissent dans des groupes nombreux, pris au hasard. »

Le même auteur dénonce ce qu'il appelle le dangereux mythe de la courbe de Gauss :

« Dans les sciences humaines, la courbe en cloche de Gauss joue un rôle considérable, parce qu'elle est l'image même de la répartition de bien des aptitudes et des qualités : les individus moyens abondent, mais les génies et les idiots, les géants et les nains sont rares. Comme les tests mesurent souvent des aptitudes, des traits de personnalité ou des performances de vastes populations, et servent à classer les individus en les comparant les uns aux autres, il est naturel que ces épreuves soient étalonnées selon la répartition gaussienne : en gros, 70 % de moyens, 13 % de bons, 13 % de médiocres, 2 % d'excellents, 2 % de très mauvais. »

Mais, l'école n'a pas, en principe, pour visée première, la sélection. Il convient donc de s'interroger sur la fatalité de la répartition gaussienne des résultats, d'autant que le traitement réservé à chaque élève dépend, dans ce modèle, de sa position relative initiale dans le groupe d'apprentissage (Crahay, 1996, 2000).

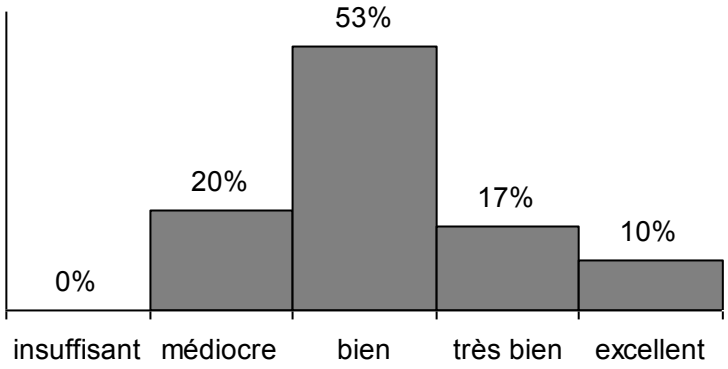
De l'intériorisation de cette distribution forcée découlent deux phénomènes particuliers:

[5] Laugier et Weinberg (1927) souscrivent à cette idée: *« En gros, les notes* [de 166 candidats à un concours universitaire dont les copies ont été jugées par deux correcteurs indépendants, expérimentés et méticuleux] *sont distribuées par chaque examinateur à peu près suivant une courbe normale "en cloche": les notes moyennes sont les plus fréquentes, les notes très bonnes ou très mauvaises sont les plus rares. C'est un signe qui atteste de la valeur de la notation, car on sait que si l'on mesure, dans un groupe homogène d'individus, un trait quelconque, - que ce soit la taille ou le poids, ou une fonction mentale au moyen d'un test, - on constate que les résultats se distribuent selon une courbe en cloche. Tout porte à croire qu'il en est de même pour les connaissances dans le groupe d'individus qui se présentent à un concours, et la confirmation en a été donnée par les applications de tests pédagogiques. »*

a) L'effet de tendance centrale :

On observe fréquemment que les notateurs de performances concentrent leurs appréciations sur les échelons du centre de l'échelle. On peut y trouver deux grands types d'explications. La première est qu'ils ont la courbe de Gauss en tête, et se figurent donc que le plus grand nombre DOIT se trouver au centre de l'échelle. Certains juges vont même jusqu'à modifier certaines des notes pour que la courbe soit plus « parfaite ». La deuxième explication est la prudence (ou la lâcheté) puisqu'en donnant une note centrale, le correcteur ne peut jamais être aux « antipodes » de la note vraiment méritée par la performance.

Rot et Butas (1959) rapportent que Gjorgjevski a invité 5 professeurs d'une même branche de l'enseignement secondaire à noter indépendamment les uns des autres 100 copies de leur discipline sur une échelle à 5 degrés (1 = *INSUFFISANT*; 2= *MEDIOCRE*; 3 = *BIEN*, 4 = *TRES BIEN* ; 5 = *EXCELLENT*). Il a ensuite extrait 15 copies qui avaient toutes reçu la note *« BIEN »* par les 5 correcteurs. Elles ont été confiées, pour nouvelle correction, à 4 autres professeurs, qui ont à nouveau distribué les 15 copies à travers les 5 catégories de notes, comme l'indique la figure ci-dessous.



b) La surprenante stabilité des taux de réussite et d'échec d'année en année.

Certains enseignants sont fiers d'une telle stabilité, preuve pour eux que « l'ordre des choses » (la courbe de Gauss et un score de passage toujours fixé au même endroit) est « respecté ». Hutmacher (1993), à Genève, a développé une version de cette théorie adaptée à l'enseignement primaire, ce qu'il appelle l'hypothèse socio-arithmétique selon laquelle les maîtres ont dans la tête le nombre « normal » d'échecs (redoublements) par classe : 2 élèves, et font en sorte que ce résultat soit observé. Ce qui débouche sur la conséquence paradoxale que plus la classe est petite (10 élèves par exemple), plus le taux d'échecs est élevé (20% pour 10 élèves, 10% pour 20 élèves).

3.2. Les biais résultant de l'interaction entre le correcteur et l'étudiant ou la copie évalué

Dans le cas des évaluations internes, menées par l'enseignant lui-même, un certain nombre d'effets indésirables peuvent résulter de la connaissance que l'enseignant a de l'élève et de l'idée qu'il se fait de ses compétences, *a priori*.

3.2.1. Effet de stéréotype ou d'inertie

Le premier type de "parasitage" de la note peut résulter d'une sorte d'effet d'inertie: le correcteur a tendance à attribuer à un étudiant une note comparable à celles que celui-ci a acquises auparavant.

« La connaissance des résultats antérieurs d'un élève – même inconnu - tend à influencer l'évaluateur. On assiste à une sorte d'imitation par contagion... Par stéréotypie, on entend une immuabilité plus ou moins accusée qui s'installe dans le jugement porté sur l'élève », comme le précise De Landsheere (1992, pp. 47-48).

Caverni, Fabre et Noizet (1975) ont mené l'étude suivante. A des professeurs de sciences de l'enseignement secondaire, ils ont demandé de noter (sur 20) chacun les 4 mêmes copies, accompagnées de « 5 notes censées avoir été obtenues précédemment par l'auteur de la copie ». Chaque série de 5 notes avait deux caractéristiques : sa moyenne (élevée = 13/20 ou faible = 7/20) et sa dispersion, exprimée par la Marge de Variation (MV), c'est-à-dire l'écart entre les notes extrêmes (MV forte = 10 ; MV faible = 2).

De Landsheere (1992, p. 47) commente : « *La moyenne exprimait le niveau moyen de l'élève, tandis que la dispersion exprimait la régularité ou l'irrégularité de ses performances. Un autre descripteur aurait (encore) pu être utilisé : la succession des notes peut marquer un progrès (ce qui était le cas ici pour toutes les copies) ou, au contraire, une régression.* »

Le tableau ci-dessous indique le résultat obtenu pour chacune des 4 copies (a, b, c et d) dans deux situations particulières : l'information sur les résultats préalables faisait apparaître une moyenne forte (13/20) et une marge de variation faible (2 points), dans le premier cas, et l'inverse (moyenne faible et marge de variation forte), dans le second cas. Comme on peut le constater, la seconde situation est plus défavorable que la première.

Copies :	a	b	c	d	Moyenne (sur les 4 copies)
Moyenne forte Marge de Variation faible	12	8,5	15,25	3	9,69
Moyenne faible Marge de Variation forte	9,75	6,5	11,75	2,75	7,69

Mais, comme le précise De Landsheere (1992, p. 48) : *« On aurait tort de croire que la stéréotypie influence uniquement les évaluations à base subjective accusée... Elle atteint des exercices aussi « objectifs » que la dictée orthographique. L'expérience suivante (inspirée de Zillig, 1967) en témoigne. Un professeur de langue maternelle fait régulièrement des dictées. Bientôt, il connaît les élèves qui réussissent habituellement le mieux et le moins bien cet exercice. Si l'on détermine la fréquence des fautes « oubliées », non perçues par le correcteur, on constate que les oublis en faveur des bons élèves sont significativement plus élevés que pour les élèves faibles. Dans le premier cas, le maître s'attend à ne pas rencontrer d'erreurs ; dans le second, il les guette. »*

Noizet et Caverni (1978, p.141) notent : *« Il est probable que les premiers indices recueillis, qu'ils soient positifs ou négatifs, vont… guider le recueil des indices… l'évaluateur cherchant davantage des indices susceptibles de confirmer ses premières inférences que des indices*

susceptibles de les remettre en question ». Et De Landsheere (1992, p. 54) poursuit : *« Bref, il semble que s'il doit faire des fautes, l'élève a intérêt à les faire dans la seconde moitié de son examen. C'est… ce qu'une expérience rapportée par Noizet et Caverni (p.142) confirme. »*

3.2.2. Effet de halo

Un autre type de "parasitage" de la note résulte de l'influence de celle-ci par des aspects non pertinents. Dans ce cas, par exemple, la note est influencée (« contaminée ») par des caractéristiques de l'étudiant comme son aspect physique, sa présentation vestimentaire, sa prononciation ou son accent, etc.

« L'effet de halo présente un caractère affectif accusé. Souvent, on surestime les réponses d'un élève de belle allure, au regard franc, à la diction agréable…Soit pour des raisons de lisibilité, soit pour des raisons nettement affectives, l'écriture peut aussi influencer le correcteur. » (De Landsheere, 1992, p. 49). Dans cet ordre d'idée, Chase (1968) a montré que la mauvaise qualité de l'écriture fait baisser le score.

Weiss (1969), de son côté, a fait l'expérience suivante (rapportée par De Landsheere, 1992, p. 50) :

Deux rédactions dactylographiées ont été soumises à 2 groupes de 46 instituteurs de 4° primaire. Au groupe 1, il dit

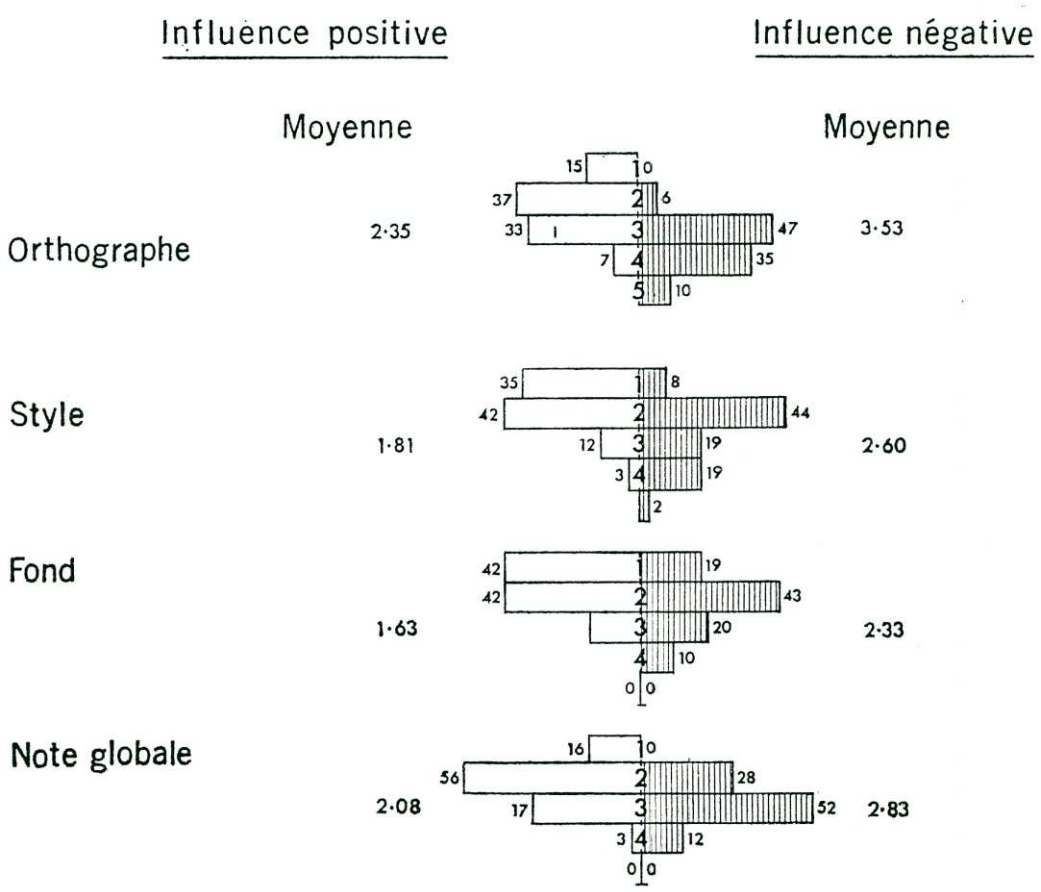
« Le travail 1 est l'œuvre d'un élève moyen qui aime lire des BD ; son père et sa mère sont employés. Le travail 2 a été fait par un enfant doué ; son père est rédacteur d'un quotidien connu ».

Pour le groupe 2, les commentaires ont été inversés. Trois aspects (orthographe, Style, Fond) devaient être jugés indépendamment, en plus d'une « note Globale », chaque fois sur une échelle à 5 niveaux (1 = TB ; 5 = insuffisant).

Comme le précise De Landsheere (1971, p. 35), dont nous reproduisons la figure inspirée des résultats de Weiss, *« Pour les quatre aspects considérés, les notes attribuées au travail pour lequel on a créé un préjugé favorable ont été significativement supérieures aux autres. Pour l'orthographe, qui semblait le plus devoir échapper à l'effet œdipien de la prédiction*[6]*, on observe qu'au travail de l'élève présenté comme doué, 16% des correcteurs accordent la note très bien et aucun la note insuffisant; si le même élève est présenté comme moyen, les correcteurs n'accordent aucun très bien, mais 11% notent insuffisant ».*

[6] Effet œdipien de la prédiction: expression due à K. Popper (1957) (Oedipus effect of prediction), en référence au personnage mythologique. D'autres synonymes, comme *effet d'anticipation de l'expérimentateur* ou encore *effet Rosenthal*, en référence aux expériences de Rosenthal et Jacobson (1971, pour la traduction française) sont également utilisés. Il s'agit, selon De Landsheere (1979, p. 104) de l'*effet que la prédiction d'un événement ou la croyance à sa venue, chez un sujet impliqué dans une situation, exerce sur la réalisation de la prédiction.* Rosenthal et Jacobson parlent, en ce qui les concerne, de *"réalisation automatique des prophéties"* ou, plus exactement, en langue anglaise, de *"Self-Fulfilling Prophecy"*, dans leur texte original de 1968. Selon De Landsheere, *« l'expression effet Rosenthal devrait être réservée au phénomène où l'anticipation de l'expérimentateur, due à une prophétie, modifie le comportement de celui-ci, de façon telle qu'il augmente la probabilité que l'événement se produise. C'est ce que Merton appelait "la prophétie qui s'exauce". »*

NOTATION DES COMPOSITIONS SOUS L'INFLUENCE D'UN PREJUGE FAVORABLE OU DEFAVORABLE [1]



Ebel (1965, p. 183) note ainsi que des contre-performances lors de tests peuvent être révélatrices d'évaluation surfaites lors de situations non standardisées.

3.3. Effets de contraste entre copies ou entre étudiants

Plusieurs effets parasites peuvent être identifiés comme relevant de l'interaction entre copies successives. On les qualifient d'*Effets* ***de contraste ou de séquence*** : la copie qui suit une copie brillante risque d'être désavantagée et inversement.

De Landsheere (1992, p. 52) décrit ce phénomène de la manière suivante: *« Les élèves rompus aux examens ont depuis longtemps découvert l'importance des contrastes : passer immédiatement après un candidat brillant se révèle défavorable ; succéder à plus faible que soi peut être avantageux, à condition que la médiocrité des réponses que l'interrogateur vient d'obtenir ne l'ait pas mis de trop méchante humeur. »*

De manière à mettre le phénomène en évidence, Bonniol (1972) a présenté une série de devoirs à corriger par deux groupes de 9 correcteurs. Ce sont les mêmes devoirs, mais ils sont présentés dans l'ordre inverse dans les deux groupes. Il observe que les différences (importantes) entre les deux groupes *« sont plutôt imputables aux deux ordres de correction qu'aux différences de critères dont les examinateurs font état »*.

A partir de cette observation, il a décidé d'introduire systématiquement après le premier tiers et après le deuxième tiers de la séquence des copies initiales des copies aux propriétés (valeur

de la note) connues : faibles ou très bonnes. Bonniol appelle ces copies des ancres. Il définit le concept d'ancre de la manière suivante: *« un stimulus privilégié qui joue comme un stimulus de référence, soit parce qu'il est présent plus fréquemment que les autres, soit parce qu'il est situé dans une position particulière, soit parce qu'il est signalé d'une manière ou d'une autre à l'attention du sujet ».*

Il appelle *Ancre Haute* une copie meilleure que les autres, et *Ancre Basse* une copie moins bonne que les autres. Pour lui est une *Ancre Lourde* la succession de trois ancres du même type.

Pour De Landsheere (1992, p. 53), *« On émet deux hypothèses : l'introduction des ancres exercera des effets de contraste, se traduisant par des déplacements dans l'échelle d'évaluation par surestimation ou sous-estimation des travaux succédant à l'ancre dans la série, et par modification de l'étendue de l'échelle utilisée. »* Il rapporte que dans sa série d'expériences, Gjorgjevski a extrait de ses 100 copies,

12 jugées *« TRES BIEN »* (échelon 4)

12 jugées *« MEDIOCRE »* (échelon 2).

Dans chacun de ces groupes de 12, il a glissé 3 copies jugées *« BONNES »* (échelon 3).

Dans le premier groupe, les trois copies ont vu leur moyenne passer de 3 à 2,4 et dans l'autre groupe, de 3 à 3,87. Ce qui confirme les deux hypothèses signalées par De Landsheere.

3.4. L'instabilité d'un même correcteur

Les effets de séquence, de contraste, etc. mais aussi des variations internes au correcteur (fatigue ou distraction momentanée, hasard..) font qu'un même correcteur peut, à des moments différents, donner des notes différentes à une même copie. Avec quelles conséquences pour les candidats, se demandera-t-on dans une perspective pratique ?

Comme le rappelle De Landsheere (1992, p. 45), *« Hartog et Rhodes (1935, p. 15) ont demandé à 14 historiens de noter une deuxième fois 15 compositions 12 à 19 mois après les avoir notées une première fois. Toute trace de correction avait été effacée. Les professeurs accordaient non seulement des points, mais indiquaient la réussite globale ou l'échec. Dans 92 cas sur 210, soit près de la moitié des cas, le verdict a été différent d'une fois à l'autre.* »

3.5. Les différences entre correcteurs

On peut mettre en évidence, en faisant corriger la même copie par plusieurs correcteurs qualifiés, des différences parfois fort importantes entre les notes attribuées à celle-ci. Ce type d'études a été mené très tôt.

Dans une expérience, rapportée par Piéron (1963, p. 123), une même composition française a été jugée par 76 professeurs de français. Voici la distribution de leurs notes (NP = Nombre de correcteurs attribuant une note donnée) :

Note	0-1	2-3	4-5	6-7	8-9	10-11	12-13
NP	1	6	20	34	10	3	2

De manière à corriger ce phénomène, Laugier et Weinberg ont appelé valeur « vraie » la moyenne d'un nombre assez grand de notations indépendantes, pensant qu'en multipliant les

correcteurs, on compensera leurs fluctuations (Piéron, 1963, p.22). Ils ont cherché à déterminer le nombre minimum d'examinateurs compétents auxquels il faudrait faire appel pour obtenir la notation méritant confiance. Dans ce but, ils ont utilisé la formule de Spearman-Brown qui a été présentée dans la partie relative à l'accroissement de la fidélité des tests en fonction de la longueur, selon la théorie classique.

Pour rappel, la formule de Spearman-Brown, qui a été décrite de manière générale pour tout allongement quelconque d'un test par un coefficient m, peut s'écrire de la manière suivante :

$$\rho_{mm} = \frac{m\rho_{11}}{1+(m-1)\rho_{11}}$$

où ρ_{11} est la fidélité du test initial

ρ_{mm} est la fidélité du test de longueur modifiée.

Au départ de cette formule, les auteurs vont considérer la fidélité inter-correcteurs comme la fidélité originale ρ_{11}. Cette fidélité inter-correcteurs est établie sur la base de la corrélation des notes remises par deux correcteurs confrontés aux mêmes copies. S'il existe plus de deux correcteurs, on calculera la corrélation moyenne au départ de toutes les corrélations calculables entre les notes transmises par chaque paire de correcteurs. La valeur ρ_{mm} sera la fidélité inter-correcteurs qui résulterait de la multiplication du nombre de correcteurs par m.

Ainsi, si 4 correcteurs fournissent une fidélité inter-correcteurs moyenne de 0,870, on obtiendra respectivement les fidélités inter-correcteurs suivantes:

0,953 pour 12 correcteurs (soit m= 3),

0,964 pour 16 correcteurs (soit m= 4),

0,982 pour 32 correcteurs (soit m= 8).

Dans le cas du doublement du nombre de correcteurs (soit 8 correcteurs et m= 2), la formule, dans le premier cas, devient en effet, après substitution:

$$\rho_{mm} = \frac{2*0,87}{1+(2-1)0,87} = 0,93$$

On peut aussi se poser le problème inverse: de combien de correcteurs devrait-on disposer pour obtenir une fidélité inter-correcteur donnée. La réponse s'obtient à partir d'une simple transformation de la formule de Spearman-Brown:

$$m = \frac{\rho_{mm}(1-\rho_{11})}{\rho_{11}(1-\rho_{mm})}$$

Ainsi, si l'on désire une fidélité inter-correcteurs d'au moins 0,90, alors que la fidélité inter-correcteurs moyenne de départ, établie sur 4 correcteurs est de 0,87, on devra multiplier le nombre de correcteurs par 1,34, ce qui impliquerait 6 correcteurs (en fait, la valeur calculée indique 5,4, mais il faut bien envisager le recours à des correcteurs entiers !). La formule, appliquée dans le cas d'une fidélité inter-correcteurs de 0,99, s'écrira de la manière suivante:

$$m = \frac{0{,}99 * (1 - 0{,}87)}{0{,}87 * (1 - 0{,}99)} = 14{,}8$$

Dans ce cas particulier, il faudra donc avoir recours à 14,8 * 4 correcteurs, soit environ 60 correcteurs (en fait, 59,2 d'après le calcul).

Se basant sur cette formule, Piéron (1969, p. 23) rapporte les résultats estimés par Laugier et Weinberg pour ce qui concerne les épreuves du baccalauréat:

« Recherchant un coefficient élevé de fidélité (0,99), et se fondant sur les moyennes des indices de corrélation obtenus pour chaque catégorie d'épreuves, ils ont trouvé que ce nombre minimum était le suivant :

Domaine	Nombre estimé de correcteurs pour obtenir une fidélité inter-correcteurs de 0,99
Composition française	78
Version latine	19
Anglais	28
Mathématique	13
Dissertation philosophique	127
Physique	16

D'autres résultats, des mêmes auteurs, sont rapportés par Agazzi (1967, p. 119): Pour les 6 mêmes domaines, 6 correcteurs ont chaque fois noté les examens de 0 à 20, une note inférieure à 10 signifiant l'échec. Le tableau suivant présente les résultats obtenus pour l'ensemble des 6 correcteurs. La première colonne indique le nombre de copies refusées par les 6 correcteurs, la dernière, le nombre de copies acceptées par ces 6 correcteurs et la colonne du centre, le nombre de copies pour lesquelles on enregistre au moins une note discordante (au moins un refus et 5 notes suffisantes ou l'inverse).

	6 notes insuffisantes	Au moins un avis discordant	6 notes suffisantes
Version latine	40%	50%	10%
Composition française	21%	70%	9%
Anglais	37%	47%	16%
Mathématique	44%	36%	20%
Philosophie	9%	81%	10%
Physique	37%	50%	13%

Piéron (1969) et De Landsheere (1971) ne manquent pas de présenter d'autres exemples encore.

Ce type de résultats a, entre autre, provoqué la remise en cause des notes chiffrées. Malheureusement, d'autres expériences ont aussi montré les limites du système d'appréciations globales du type "Très bien, Bien, Satisfaisant, Faible, Insuffisant", comme nous allons le voir.

On n'a pas manqué de penser que des expressions verbales (d'ailleurs en nombre plus limité que 21 notes possibles) augmenteraient la concordance inter-correcteurs. D'où l'adoption par des systèmes scolaires entiers (la Communauté française de Belgique, par exemple) de ce type d'échelles.

Reuchlin (1958, 1968) avertit cependant du danger :

« ...l'instituteur, certainement, connaît mieux que personne les points du programme qui sont acquis ou non par chacun de ses élèves. Ce qu'il ignore, c'est la gravité qui s'attache à chaque faiblesse, à chaque lacune, lorsqu'on la considère non plus au sein d'une classe qui peut être « forte » ou « faible », mais par rapport à l'ensemble du pays. De là, les divergences d'appréciation mises en lumière par l'enquête. »

Ces divergences sont illustrées par les 4 courbes ci-dessous. Elles sont issues d'une enquête nationale française, menée en 1958, ici sur le calcul au « cours moyen 2° année », ce que l'on appelle la 5° primaire en Belgique[7].

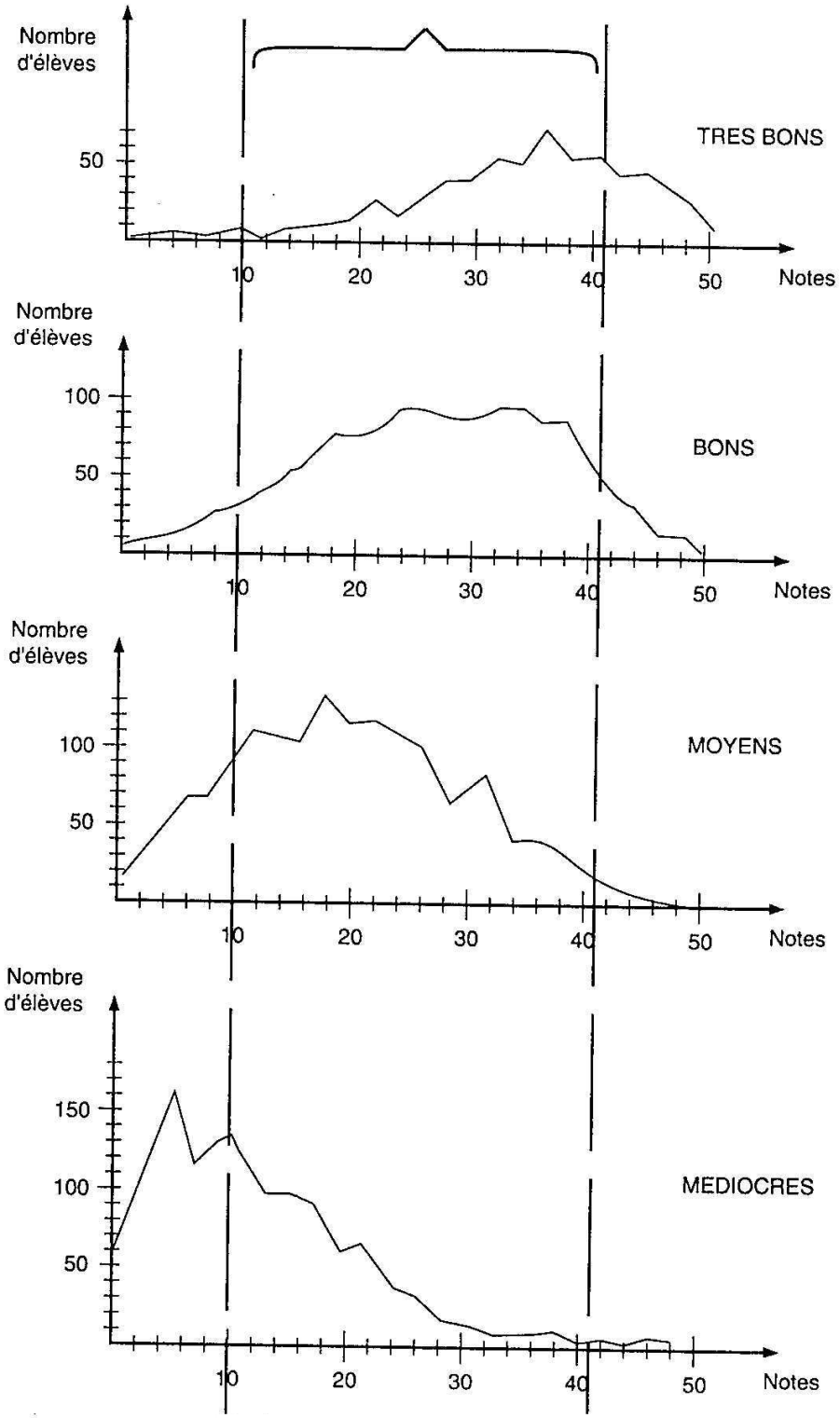
Les instituteurs avaient été invités à attribuer à chaque élève un des 4 adjectifs suivants pour caractériser son niveau en calcul :

TRES BON, BON, MOYEN, MEDIOCRE.

De cette manière, 654 élèves furent jugés *TRES BONS*, 1303 *BONS*, 1551 *MOYENS*, 1300 *MEDIOCRES*. La catégorie "moyen" est celle qui rassembla le plus grand nombre d'élèves. Ces élèves ont par ailleurs subi un test de calcul noté « objectivement » de 0 à 50. La figure ci-dessous reproduit, pour chacune des "catégories d'élèves" résultant de l'avis des maîtres, la distribution des notes au test.

Ces quatre distributions se recouvrent largement : dans la zone de notes qui va de 10 à 40, le même niveau de performance au test peut malheureusement correspondre à n'importe laquelle des 4 notes verbales globales.

[7] Il s'agit de la dernière année de l'enseignement primaire français, celui-ci ne comportant que 5 années.



4. En guise de conclusion

Nous l'avons précisé avant d'entamer ce chapitre, la docimologie critique a permis d'attirer l'attention des correcteurs et des enseignants sur la nécessité d'un soin particulier quand aux différents parasitages possibles de la notation. Les expériences nombreuses et déjà anciennes

qui sont mentionnées, et qui ne constituent qu'un petit échantillon de ce qui a pu être réalisé entre les années 1920 et les années 1970, ne doivent pas conduire au rejet absolu de la notation subjective. C'est impossible. Ces résultats ne doivent sans doute pas plus conduire au rejet de toute forme d'évaluation en dehors de l'usage de questions à réponse fermée (vrai/faux, QCM), mais faire réfléchir à des formules efficaces - la méthode d'examen doit rester praticable dans des conditions normales - et justes. Ce second critère est essentiel car il s'agit d'apprécier les compétences de sujets humains. Les examens décident de plus en plus, quand il ne s'agit pas de concours, du sort de personnes. Il convient d'y être attentif. Et si cette attention n'est pas nécessairement spontanée chez tous les examinateurs, le risque est grand de voir intervenir de plus en plus d'autres acteurs dans la sphère scolaire à l'occasion de l'évaluation. On constate en effet que la judiciarisation[8] et la juridiciarisation[9] sont deux menaces importantes qui pèsent sur la liberté énorme, pour ne pas dire totale, qui avait prévalu jusqu'il y a peu dans le domaine de l'évaluation scolaire. Pour s'en convaincre, il suffit de se reporter au dossier d'information[10] préparé pour la journée organisée conjointement par l'*Association pour le Développement des Méthodologie d'Evaluation en Education* et l'*Association internationale de Pédagogie universitaire,* le 4 décembre 2001 à Liège. On peut notamment consulter l'article de V. De Landsheere sur "la responsabilité civile découlant de l'enseignement dispensé" et qui est paru dans le *Journal des procès* du 30 décembre 1988 (n° 141, pp. 10-13). Ce dernier attire l'attention sur l'importance du phénomène de judiciarisation aux Etats-Unis, il y a plus de quinze ans, et peut préfigurer ce qui pourrait se produire de plus en plus chez nous, si on n'apporte pas un soin particulier aux évaluations.

Bibliographie

Agazzi, A. (1967). *Les aspects pédagogiques des examens.* Strasbourg: Conseil de l'Europe.

Barrere, A. (2000). Sociologie du travail enseignant. *L'Année sociologique*, 50(2), 469-492.

Baudelot, C., Establet, R. (1971). *L'école capitaliste en France*. Paris: Maspero.

Bonniol, J.-J., Piolat, M. (1971). Comparaison des effets d'ancrage obtenus dans une tâche d'évaluation. Expérience de multi-correction en mathématique et en anglais. in *Actes du XVII^e Congrès international de psychologie appliquée*, 8 , 1179-1189.

Bonniol, J.-J. (1965). Les divergences de notation tenant aux effets d'ordre de la correction. *Cahiers de Psychologie*, 8 , 181-188.

Boudon, R. (1973). *L'inégalité des chances*. Paris: Armand Collin.

Bourdieu, P., Passeron, J.-C. (1964). *Les héritiers*. Paris: Les éditions de Minuit.

Bourdieu, P., Passeron, J.-C. (1970). *La reproduction. Eléments pour une théorie du système d'enseignement.* Paris: Les éditions de Minuit.

Bressoux, P. (1994). Note de synthèse : Les recherches sur les effets-écoles et les effets-maîtres. *Revue française de pédagogie*, n° 108, 91-137.

Bressoux, P. (1995). Les effets du contexte scolaire sur les acquisitions des élèves : effet-école et effets-classes en lecture. *Revue française de sociologie*, XXXVI, 273-294.

Coleman, J. S. et al. (1966). *Report on Equality of Educational Opportunity (EEOR)*. U.S Government Printing Office for Department of Health, Education and Welfare.

[8] Intervention du pouvoir judiciaire dans le règlement de litiges.

[9] Intervention du droit dans la définition des différentes activités.

[10] On peut consulter la liste des articles qui le composent à l'adresse: http://www.ulg.ac.be/pedaexpe/judi/biblio.pdf

Caverni, J.-P., Fabre, J.-M., Noizet, G. (1975). Dépendance des évaluations scolaires par rapport à des évaluations antérieures : études en situation simulée. *Le Travail Humain*, 38(2), 213-222.

Chevallard, Y. (1991). Vers une analyse didactique des faits d'évaluation, in J.-M. De Ketele (Ed.), *L'évaluation : approche descriptive ou prescriptive*. Bruxelles: De Boeck Université.

Crahay, M. (1996). *Peur-on lutter contre l'échec scolaire ?* Bruxelles: De Boeck Université.

Crahay, M. (2000).*L'école peut-elle être juste et efficace ? De l'égalité des chances à l'égalité des acquis.* Bruxelles: De Boeck Université.

Dardenne, B. (1999). *Psychologie sociale*. Liège: Université de Liège.

De Landsheere, G. (1971, 1992). *Evaluation continue et examens. Précis de Docimologie.* Bruxelles: Editions Labor et Paris: Fernand Nathan.

Demeuse, M (2002). *Analyse critique des fondements de l'attribution des moyens destinés à la politique de discrimination positive en matière d'enseignement en Communauté française de Belgique* (Thèse doctorale). Liège: Université de Liège.

Dubet, F., Martuccelli, D. (1996). *A l'école. Sociologie de l'expérience scolaire*. Paris: Seuil, coll. « L'épreuve des faits ».

Duru-Bellat, M. (1995). Note de synthèse : Filles et garçons à l'école, approches sociologiques et psycho-sociales. *Revue française de pédagogie*, 110 , 75-109.

Duru-Bellat, M., Mingat, A. (1993). *Pour une approche analytique du système éducatif*. Paris: Presses Universitaires de France.

Felouzis, G. (1993). Interactions en classe et réussite scolaire. Une analyse des différences filles-garçons. *Revue française de sociologie*, XXXIV, 199-222.

Festinger, L. (1957). *A theory of cognitive dissonance,* Stanford, Stanford University Press.

Good, T. (1987). Two decades of research on teacher expectations : Findings and future directions. *Journal of Teacher Education*, 24, 32-47.

Grisay, A. (1984). Les mirages de l'évaluation scolaire. Rendements en français, notes et échecs à l'école primaire ? *Revue de la Direction Générale de l'Organisation des Etudes*, 1984, XIX, 5, pp. 29-42.

Grisay, A. (1991). Que peut-on prescrire en matière d'éducation-bilan ? in J.-M. De Ketele (Ed.), *L'évaluation : approche descriptive ou prescriptive*. Bruxelles: De Boeck Université.

Hutmacher, W. (1993). Quand la réalité résiste à la lutte contre l'échec scolaire. Analyse du redoublement dans l'enseignement primaire genevois. Genève: Service de la Recherche sociologique, *Cahier n°36.*

Jencks, C. (1979). *L'inégalité : influence de la famille et de l'école en Amérique*. Paris: Presses universitaires de France.

Laugier, H., Pieron, H., Pieron, H, Toulouse, E., Weinberg, D. (1934). *Etudes docimologiques sur le perfectionnement des examens et concours.* Paris: Conservatoire national des arts et métiers, Publications du *Travail humain*, Série A, n°3.

Laugier, H., Weinberg, D. (1927). Les facteurs subjectifs dans les notes d'examen. *Année Psychologique, XXVIII,* 236-244.

Laugier, H., Weinberg, D. (1936). *Commission française pour l'enquête Carnegie sur les examens et concours. La correction des épreuves écrites au baccalauréat*. Paris: Maison du livre.

Leyens, J.-Ph., Yzerbit, V. (1997). *Psychologie sociale*. Liège: Mardaga.

Merle, P. (1996). *L'évaluation des élèves. Enquête sur le jugement professoral*. Paris: Presses Universitaires de France.

Merle, P. (1998). *Sociologie de l'évaluation scolaire*. Paris: Presses Universitaires de France.

Nicaise, J. (2001). *Pratiques, sens et sens pratique au cœur des évolutions institutionnelles. Les instituteurs de sixième primaire et le jugement professoral*. Liège: Université de Liège (mémoire de licence non publié).

Perrenoud, P. (1995). *La fabrication de l'excellence scolaire : du curriculum aux pratiques d'évaluation*. Genève: Droz.

Perrenoud, P. (1998). *L'évaluation des élèves. De la fabrication de l'excellence à la régulation des apprentissages. Entre deux logiques*. Bruxelles: De Boeck Université.

Pieron, H. (1963). *Examens et docimologie.* Paris: Presses universitaires de France.

Pourtois, J-P., Bonacina, R., Delbecq, A., Segard, M. (1978). Le niveau d'expectation de l'examinateur est-il influencé par l'appartenance sociale de l'enfant ? *Revue française de pédagogie*, 44 , 34-37.

Rosenthal, R. A., Jacobson, L. (1971). *Pygmalion à l'école*. Tournai: Casterman[11].

Rot, N., Butas, Z. (1959). Les distributions des notes scolaires comparées eux distributions des résultats obtenus aux tests de connaissances. *Le travail humain, XXII,* 1-2.

Weber, M. (1971). *Economie et société*. Paris: Plon.

[11] Traduit de l'américain. Titre original: *Pygmalion in the Classroom. Teacher Expectation and Pupil's Intellectual Development* (1968).