Méthodes psychométriques en qualité de vie

Christophe Lalanne
EA 7334 REMES
Unité de Méthodologie des critères d'évaluation
Université Paris-Diderot, Sorbonne Paris-Cité

Fidélité de mesure

— Consistance interne d'une échelle
— Accord inter-juges (cas binaire ou non)
— Corrélation intra-classe

Comment évaluer la fidélité de mesure

Une formulation alternative consiste à se demander quelles sont les sources potentielles de variation des scores, et donc comment les mesurer et quel est leur impact lorsque l'on infère des résultats observés sur un échantillon à la population ? Des mesures collectées plusieurs fois à partir d'un même instrument peuvent survenir de plusieurs manière[1] :

- évaluation répétée de plusieurs sujets par le même évaluateur ;
- évaluation alternative d'un même individu par plusieurs évaluateurs ;
- administration répétée d'un même questionnaire ou de formes parallèles ;
- utilisation de différentes sous-échelles d'un même questionnaire.

1. G Dunn. *Statistics in Psychiatry*. Hodder Arnold, 2000.

Outils statistiques

La fidélité ou précision de mesure peut être quantifiée à l'aide de différentes techniques :

- **décomposition (linéaire) des composantes de variance en TCT** ;
- modèles d'équations structurelles ;
- modèles de réponse à l'item.

Fidélité de mesure ≠ significativité

La **significativité statistique** est utilisée pour évaluer la probabilité ou la vraisemblance de résultats observés sur un échantillon en référence à un modèle de population sous l'hypothèse nulle ; la **significativité pratique ou clinique** reflète le degré de divergence des résultats observés avec l'hypothèse nulle (tel que mesuré par une mesure de taille d'effet) – sous laquelle on ne distingue pas les patients des sujets contrôles. [2]

Mais ces deux concepts supposent que les scores sur lesquelles les conclusions reposent sont des indicateurs corrects et précis de la performance ou de l'état mesuré chez l'individu.

2. B Thompson, éd. *Score Reliability. Contemporary Thinking on Reliability issues*. Sage Publications, 2003.

It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees (Feldt & Brennan, 1989). Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric.

Wilkinson & APA Task Force [3]

3. L Wilkinson & APA Task Force on Statistical Inference. ≪ Statistical methods in psychology journals : Guidelines and explanations ≫. In : *American Psychologist* 54 (1999). reprint available through the APA Home Page : `http://www.apa.org/journals/amp/amp548594.html`, p. 594–604.

Modèle de mesure

▷ 01a-scores.pdf

Pour un individu i évalué sur une seule occasion, son score x_i peut être exprimé comme suit :

$$x_i = \tau_i + \varepsilon_i \qquad \varepsilon_i \sim \mathcal{N}(0; \sigma_e^2),$$

d'où l'on en déduit naturellement que $\mathbb{E}(X) = T$ (« par construction »).

Si l'on suppose que T et E sont indépendants, on a également

$$\mathbb{V}(X) = \mathbb{V}(T) + \mathbb{V}(E)$$

On peut définir le **coefficient de fidélité** de la manière suivante :

$$\begin{aligned} R_X &= \frac{\mathbb{V}(T)}{\mathbb{V}(X)} \\ &= \frac{\mathbb{V}(T)}{\mathbb{V}(T) + \mathbb{V}(E)}. \end{aligned} \tag{1}$$

Il s'agit d'une variable aléatoire donc ce n'est pas une propriété fixe d'un instrument de mesure. La racine carré de ce coefficient est appelée **erreur standard de mesure** (SEM).

Extension simple de ce modèle de mesure

Supposons que les évaluations ne dépendent pas seulement du score vrai des individus mais également de l'évaluateur (les effets étant supposés indépendants).

Si tous les sujets sont évalués par le même évaluateur, R_X se calcule tel que défini en (1). Si, au contraire, les individus sont évalués par des évaluateurs choisis aléatoirement , alors

$$R'_X = \frac{\mathbb{V}(T)}{\mathbb{V}(T) + \mathbb{V}(I) + \mathbb{V}(E)}, \tag{2}$$

et $R_X > R'_X$.

Le cas de deux instruments

Supposons que nous disposons d'une série de données appariées collectées à partir de deux instruments de mesure, $\mathcal{I}_X$ et $\mathcal{I}_Y$, pour lesquels les scores sont construits selon le même schéma :

$$\begin{aligned} X &= T_X + E_X \\ Y &= T_Y + E_Y \end{aligned}$$

Quelle est la corrélation entre X et Y ?

Les deux séries de mesure sont des réalisations des variables alétoires X et Y, mais la précision des instruments de mesure entre également en jeu.

Corrélation atténuée

La corrélation (liénaire) entre X et Y est donnée par $\rho_{XY} = \frac{\mathsf{cov}(T_X,T_Y)}{\sqrt{\mathbb{V}(T_X)\mathbb{V}(T_Y)}}$.
Un bon estimateur peut être consrtuit comme suit :

$$\begin{aligned} \hat{\rho}_{XY} &= \frac{\mathsf{cov}(X,Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}} \\ &= \frac{\mathsf{cov}(T_X+E_X,T_Y+E_Y)}{\sqrt{\mathbb{V}(T_X+E_X)\mathbb{V}(T_Y+E_Y)}} = \rho_{XY}\sqrt{R_X R_Y} \end{aligned} \tag{3}$$

La corrélation entre les données observées est atténuée par la précision de chacun des instruments de mesure.

Il en va de même dans le cas de la régression linéaire : si l'on souhaite prédire Y à partir de X à l'aide d'un modèle de régression simple, $\mathbb{E}(T_Y|T_X) = \beta_0 + \beta_1 T_X$, la pente serait estimée par $\beta_1 = \text{cov}(T_X, T_Y)/\mathbb{V}(T_X)$. En tenant compte de la précision de la mesure X, R_X, on considèrera comme estimateur :

$$\begin{aligned} \hat{\beta}_1 &= \frac{\text{cov}(X,Y)}{\mathbb{V}(X)} \\ &= \frac{R_X \text{cov}(T_X,T_Y)}{\mathbb{V}(T_X)} = R_X \beta_1 \end{aligned} \tag{4}$$

La pente de la droite de régression est atténuée (c.a.d. ramenée vers 0).

Consistance interne d'un instrument

La consistance interne d'un questionnaire permet de résumer le degré d'homogénéité des scores aux items (variance partagée) en relation avec le score total.

La consistance interne est considérée comme une borne basse de la fidélité de mesure, et elle est souvent mesurée à l'aide du **coefficient alpha de Cronbach**, qui est une mesure dépendant de l'échantillon. Lorsque $\alpha = 0{,}70$, l'erreur de mesure vaut environ 0,55 $\times$ l'écart-type du score total.

Un tel indicateur suppose l'absence de corrélation entre les erreurs de mesure, une contribution équivalente des items dans la détermination du score total, et l'unidimensionalité de l'instrument de mesure.

Indice KR-20 et alpha de Cronbach

Dans le cas des items dichotomiques, une mesure de consistance est le coefficient de Kuder-Richardson (KR-20) : [4]

$$\text{KR-20} = \frac{K}{K-1}\left[1 - \frac{\sum_k p_k q_k}{\sigma_t^2}\right] \tag{5}$$

où K est le nombre d'items, σ_t^2 la variance des scores totaux, p_k et q_k désignent les proportions d'individus avec un score à 1 ou 0 à l'item k.

Dans le cas des items polytomiques, on remplacera $p_k q_k$ par σ_k^2, la variance des scores à l'item k. [5]

4. G F Kuder et M W Richardson. « The theory of the estimation of test reliability ». In : *Psychometrika* 2 (1937), p. 151–160.

5. L J Cronbach. « Coefficient alpha and the internal structure of tests ». In : *Psychometrika* 16 (1951), p. 297–334.

Intérêts et limites

Coefficients are a crude device that does not bring to the surface many subtleties implied by variance components. In particular, the interpretations being made in current assessments are best evaluated through use of a standard error of measurement. [6]

— L'alpha de Cronbach n'est pas une mesure de l'unidimensionalité [7] ;
— L'alpha de Cronbach, comme toute estimation de fidélité de mesure des scores, dépend de l'échantillon ; il dépend également du nombre d'items.

6. LJ Cronbach et RJ Shavelson. « My current thoughts on coefficient alpha and successor procedures ». In : *Educational and Psychological Measurement* 64.3 (2004), p. 391–418.

7. J E Danes et O K Mann. « Unidimensional measurement and structural equation models with latent variables ». In : *Journal of Business Research* 12 (1984), p. 337–352.

Pour une taille d'échantillon fixe ($N = 300$), l'alpha de Cronbach augmente de manière monotone avec le nombre d'items inclus dans l'échelle ou le degré d'intercorrélation (ρ) entre les items. Avec 30 items et $\rho = 0.350$, on peut obtenir $\alpha = 0.943$ ($\alpha = 0.910$ avec 20 items).

Alternatives au coefficient α :

— Guttman Lambda 6 (part de variance de chaque item expliquée par la régression de l'ensemble des autres items, ou plus exactement la variance des erreurs) ;
— Revelle beta, etc.

Illustration

Cas d'un instrument avec 4 items, ayant la même erreur de mesure $(0{,}36)$:

```
library(psych)
set.seed(101)
x <- sim.congeneric(loads = rep(0.8, 4),                    ❶ ❷
                    N = 100, short = FALSE)
```

```
> alpha(x$observed)

Reliability analysis
Call: alpha(x = x$observed)

  raw_alpha std.alpha G6(smc) average_r S/N  ase   mean  sd
      0.88      0.88    0.84      0.64 7.1 0.02 -0.071 0.8
```

```
 lower alpha upper     95% confidence boundaries
0.84 0.88 0.92

 Reliability if an item is dropped:
   raw_alpha std.alpha G6(smc) average_r S/N alpha se
V1      0.82      0.82    0.76      0.60 4.6    0.031
V2      0.84      0.84    0.78      0.63 5.1    0.028
V3      0.85      0.85    0.80      0.66 5.9    0.025
V4      0.86      0.86    0.80      0.67 6.0    0.025
```

Fidélité inter-cotateur

Le coefficient Kappa de Cohen est utilisé pour évaluer la concordance dans le cas de deux évaluations (supposées indépendantes) permettant de classer un individu selon un critère binaire[8] (p.ex., radio pulmonaire pour la tuberculose) ou d'attribuer un score numérique à un individu[9] (p.ex., sévérité du trouble psychiatrique).

$$\kappa = \frac{\overbrace{\sum_i \pi_{ii}}^{\text{raw agreement}} - \sum_i \pi_{i\bullet}\pi_{\bullet i}}{1 - \underbrace{\sum_i \pi_{i\bullet}\pi_{\bullet i}}_{\text{random ratings}}} \tag{6}$$

8. J Cohen. « A coefficient of agreement for nominal scales ». In : *Educational and Psychological Measurement* 20 (1960), p. 37–46.

9. J Cohen. « Weighted kappa : Nominal scale agreement with provision for scales disagreement of partial credit ». In : *Psychological Bulletin* 70 (1968), p. 213–220.

Le coefficient κ est asymptotiquement équivalent à l'ICC estimé à partir d'une ANOVA à deux effets aléatoires, même si les tests de significativité et les intervalles de confiance ne sont pas adaptés au cas binaire.

Il existe des extensions pour le cas où il y a plus de 2 cotateurs,[10] ainsi que des règles d'interprétation des valeurs de κ.[11]

10. Sidney Siegel et Jr N John Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 1988.

11. J L Fleiss. *Statistical Methods for Rates and Proportions*. New York : Wiley, 1981.

Concordance inter-juges et études diagnostiques

Exemple d'un diagnostique (malade/pas malade) fourni par deux psychiatres[12] :

		A		
		Case	Non-case	Total
B	Case	p_{11}	p_{12}	$p_{1\bullet}$
	Non-case	p_{21}	p_{22}	$p_{2\bullet}$
	Total	$p_{\bullet 1}$	$p_{\bullet 2}$	1

La concordance brute vaut $P_o = p_{11} + p_{22}$ et le taux attendu sous l'hypothèse d'un jugement au hasard $P_c = p_{1\bullet}p_{\bullet 1} + p_{2\bullet}p_{\bullet 2}$.
On a donc : $\kappa = (P_o - P_c)/(1 - P_c)$.

12. Graham Dunn. *Statistics in Psychiatry*. Hodder Arnold, 2000.

Illustration

Cas de l'évaluation de 118 tumeurs sur une échelle en 5 points[13] :

```
data(pathologist.dat, package = "exactLoglinTest")
aa <- xtabs(y ~ A + B, data = pathologist.dat)
```

```
> aa
   B
A     1  2  3  4  5
  1 22  5  0  0  0
  2  2  7  2  1  0
  3  2 14 36 14  3
  4  0  0  0  7  0
  5  0  0  0  0  3
```

13. Alan AGRESTI. *Categorical Data Analysis*. Wiley-Interscience, 1990.

```
> sum(diag(aa))/sum(aa)                                                    ❶
[1] 0.6355932
> cohen.kappa(aa)
Call: cohen.kappa1(x = x, w = w, n.obs = n.obs, alpha = alpha)

Cohen Kappa and Weighted Kappa correlation coefficients and confidence boundarie
                 lower estimate upper
unweighted kappa  0.39     0.50  0.61
weighted kappa    0.70     0.78  0.86                                      ❷

 Number of subjects = 118
```

Autres mesures d'association

Les coefficients de **corrélation tétrachorique** (cas binaire) et **polychorique** (cas ordinal) peuvent être utilisés comme mesure de concordance entre juges.

En effet, ils permettent

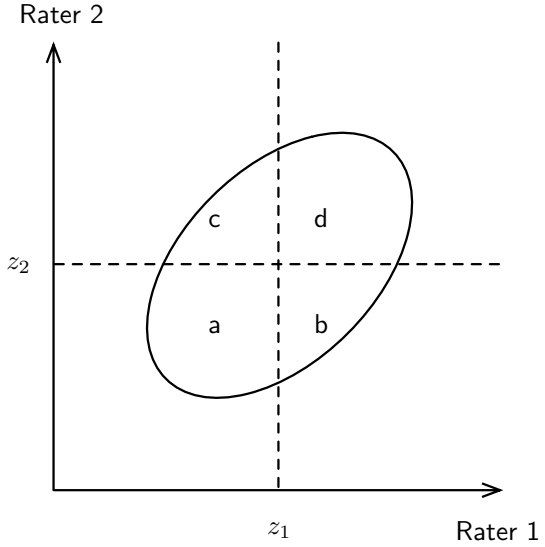
— d'estimer l'association entre les deux évaluations si celles-ci étaient réalisées sur une échelle continue ;

— de tester l'homogénéité marginale entre les juges.

Dans ce cas, on se rapproche des modèles en traits latents,[14] avec des hypothèses plus souples sur les distributions.

14. J S Uebersax. *The tetrachoric and polychoric correlation coefficients*. Statistical Methods for Rater Agreement web site. 2006.

Cas de la corrélation tétrachorique



La figure de gauche représente graphiquement le tableau croisé de deux évaluations :

		Rater 1		
		−	+	
Rater 2	−	a	b	$a+b$
	+	c	d	$c+d$
		$a+c$	$b+d$	1

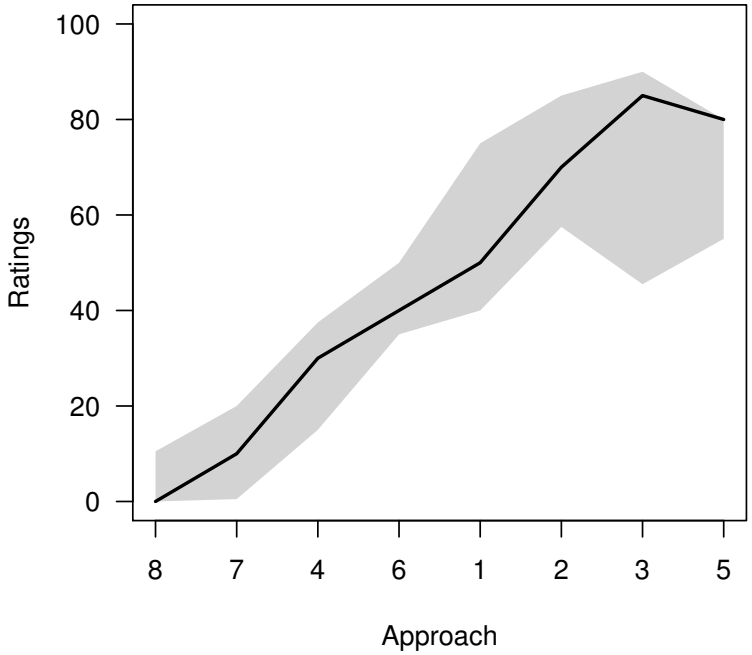
Ici $a = \Pr(Y_1 < z_1 \text{ et } Y_2 < z_2)$ est la fonction de répartition d'une distribution binormale, c.a.d. $\int_{-\infty}^{z_2}\int_{-\infty}^{z_1} \Phi(x,y,r)dxdy$, où $\Phi(\bullet)$ est la fonction de densité d'une loi binormale, avec $\Phi(z_1) = a + c$ et $\Phi(z_2) = a + b$ (valeurs seuil).

Illustration : plusieurs cotateurs

Cas de l'évaluation de 7 experts sur une échelle en 100 points portant sur 8 approches permettant de résumer la qualité de vie. [15]

Approach	Rater							Mean	SD
	1	2	3	4	5	6	7		
1	90	00	50	95	30	60	50	53.57	33.00
2	90	00	70	100	60	55	80	65.00	32.79
3	90	51	40	90	25	100	85	68.71	29.44
4	30	52	05	30	–	10	40	27.93	17.78
5	80	50	80	60	80	50	100	71.43	18.64
6	30	100	05	50	50	40	40	45.00	28.72
7	20	70	00	20	10	00	01	17.29	24.86
8	20	90	00	00	00	00	01	16.57	33.18

15. Ron D Hays et Dennis Revicki. « Reliability and validity (including responsiveness) ». In : *Assessing Quality of Life in Clinical Trials*. Sous la dir. de Peter Fayers et Ron Hays. Oxford University Press, 2005. Chap. 1.3, p. 25–39.



Médiane ± IQR (figure de gauche).
Les approches considérées comme les plus adéquates (scores plus élevés) sont associées à une plus grande variabilité entre juges même si la corrélation de rang (Spearman) entre la moyenne et l'écart-type n'est pas significative ($\rho = -0.167, p = 0.703$).

Analyse des composantes de variance

Les commandes R suivantes permettent d'estimer un modèle à un facteur (effet fixe) ❶, un modèle à deux effets ❷, et un modèle mixte incluant deux effets ❹.

```
data(hays05)
require(lme4)
summary(aov(score~approach, hays05))                    ❶
summary(aov(score~rater+approach, hays05))              ❷
summary(aov(score~approach*rater, hays05))              ❸
summary(lmer(score~approach+(1|rater), hays05))         ❹
```

Le modèle ❷ est largement utilisé en pratique et permet d'estimer l'ICC dit de consistence ou de condordance. [16]

16. P E Shrout et J L Fleiss. « Intraclass correlation : Uses in assessing rater reliability ». In :

Source	Df	MS	
Approach	7	3597.8	❶
Within	47	792.6	❶
Rater	6	844.7	❸
Approach×Rater	41	785.0	❸
Total	54		

Model	Reliability	ICC
❶	$\frac{\mathrm{MS}_B - \mathrm{MS}_W}{\mathrm{MS}_B}$	$\frac{\mathrm{MS}_B - \mathrm{MS}_W}{\mathrm{MS}_B + (K-1)\mathrm{MS}_W}$
❷	$\frac{\mathrm{MS}_B - \mathrm{MS}_E}{\mathrm{MS}_B}$	$\frac{\mathrm{MS}_B - \mathrm{MS}_E}{\mathrm{MS}_B + (K-1)\mathrm{MS}_E}$
❸	$\frac{N(\mathrm{MS}_B - \mathrm{MS}_E)}{N\mathrm{MS}_B + \mathrm{MS}_J - \mathrm{MS}_E}$	$\frac{\mathrm{MS}_B - \mathrm{MS}_E}{\mathrm{MS}_B + (K-1)\mathrm{MS}_E + K(\mathrm{MS}_J - \mathrm{MS}_E)/N}$

MS, mean square for between-subject (B), within-subject (W) effects; K, number of replications; N, number of ratees.

Psychological Bulletin 86 (1979), p. 420–428.

Fichier de données et scripts R disponibles à l'adresse suivante :
`https://bitbucket.org/chlalanne/eespe11`

– Typeset with FoilTEX (version 2), Revision e967a78