LA FIDELITE[1] DES MESURES

Marc Demeuse et Georges Henry

1. Introduction

Ce chapitre traite des problèmes de fidélité, tels qu'on les rencontre dans la pratique, aussi bien dans des situations de test qu'à travers d'autres mesures psychologiques. Bien que nous nous basions sur les théories fondamentales examinées précédemment, les mesures recueillies, au moyen de tests, ne répondent pas, le plus souvent, à toutes les exigences que nous avons énoncées. Pour cette raison, nous aurons à redéfinir opérationnellement certains concepts, notamment celui de variance vraie.

2. La fidélité des mesures

2.1. Grandes méthodes d'estimation de la fidélité

On distingue traditionnellement trois grandes familles de méthodes, qui essayent toutes de répondre à la même question " Quelle est la corrélation du test avec lui-même ? " Nous avons vu précédemment que cette corrélation du test avec lui-même correspond à la définition de la fidélité. Ces trois méthodes sont les suivantes :

(a) Méthodes basées sur la consistance interne

(b) Méthodes basées sur le test-retest.

(c) Méthodes mixtes, dites des " formes parallèles "

Ces trois méthodes ont en commun l'estimation de ρ_{tt}, fidélité du test. Elles impliquent des techniques de calcul différentes.

2.1.1. Méthodes basées sur la consistance interne

Les méthodes basées sur le ***postulat de consistance interne*** prennent comme pré-supposé de base l'une des deux idées suivantes :

> (a) tous les items du test mesurent le même chose (comme dans le cas du calcul des coefficient ***KR 20*** ou ***KR21***)
>
> (b) différentes parties du test mesurent la même chose (voir ***α de Cronbach*** et méthodes basées sur les scores obtenus aux items pairs-impairs).

2.1.2. Méthodes basées sur le test-retest

Cette famille de méthodes ne postule nullement l'existence d'une consistance interne. En fait, les différentes parties du test pourraient, à la limite, avoir une inter-corrélation nulle et, cependant, la corrélation entre une première et une seconde administration du test (fidélité test-retest) pourrait être élevée. Le concept-clé est ici celui de la stabilité dans le temps. Une corrélation élevée signifie donc que les individus demeurent plutôt stables à travers le temps et qu'ils obtiennent un score total identique ou très proche lors de plusieurs passation

[1] En anglais, généralement " ***reliability*** " mais selon la méthode adoptée pour estimer la fidélité, on trouve également " ***stability*** ", " ***consitency*** " ou " ***accuracy*** ".

consécutives, s'ils n'ont pas suivi un enseignement en rapport avec l'objet du test ou eu l'occasion de s'y entraîner. Un coefficient bas signifie, au contraire, qu'il y a fluctuation du score total, et donc de l'estimation de la compétence, à travers le temps. Le fait de représenter un même test à deux occasions pose cependant problème, surtout si le délai est court et les items suffisamment caractéristiques pour permettre une mémorisation partielle, susceptible de favoriser une meilleure réussite à la seconde occasion (***effet de testing***).

2.1.3. Méthodes mixtes (dites " des formes parallèles ")

Par définition, ces méthodes mixtes présentent des analogies avec les deux méthodes que nous venons de décrire. Ici, le résultat obtenu informe sur l'équivalence du contenu psychologique mesuré au moyen de deux formes parallèles[2]. Si ces formes ont été administrées endéans un intervalle de temps court, on mesure quelque chose de proche de la consistance interne, mais on risque aussi de voir apparaître un ***effet de testing***. Dans le cas contraire, si on présente les deux formes parallèles à une certaine distance temporelle, on mesure tout à la fois la consistance interne et la stabilité.

Il faut, en outre, être attentif au fait que, si la nature de la variance vraie diffère entre formes, dit autrement, si les deux formes ne mesurent pas exactement la même chose, la fidélité de chacune d'elles sera sous-estimée. A l'opposé, si les variances d'erreur ne sont pas indépendantes, il y aura surestimation de la fidélité. Cette surestimation résultera du biais systématique qui entache les deux formes.

2.2. Contribution des variances vraie et d'erreur dans les différentes méthodes

On comprend aisément, si on se réfère aux considérations qui précèdent, que ces trois types d'estimation de la fidélité peuvent conduire à des coefficients de fidélité très différents. En effet, ce ne sont pas les mêmes composantes qui interviennent dans la définition de la variance vraie et de la variance d'erreur. On peut identifier les différentes sources de variances vraie et d'erreur.

2 Pour qu'on puisse parler de formes parallèles, voire de formes équivalentes, il faut que celles-ci aient même variance vraie et qu'il n'y ait pas de recouvrement des variances d'erreur.

Tableau 1 – *Variance vraie et variance d'erreur dans trois méthodes d'estimation de la fidélité.*

	Variance vraie	Variance d'erreur
Consistance interne	Covariance entre items ou groupes d'items à l'intérieur d'un même test	Les items (ou groupes d'items) ne mesurent pas la même chose
Test-retest	Covariance entre les résultats du test présenté à deux occasions (test et retest)	A deux occasions, des résultats supposés identiques diffèrent en raison de conditions extérieures (fatigue différente = aléatoire car variable d'un sujet à l'autre / effet de testing = systématique, la mémorisation jouant plus ou moins fortement selon la nature du test
Formes parallèles	Covariance entre les deux formes à deux moments différents	A deux occasions, les résultats diffèrent (cf. test-retest) Les résultats aux deux formes diffèrent (cf. différences à l'intérieur d'un même test, comme dans l'étude de la consistance interne)

En outre, quelle que soit la méthode utilisée, on doit être attentif à l'existence d'un certain nombre de problèmes ou biais possibles.

- ***Expériences préalables des situations de testing*** : Plus il y a de différences entre individus, plus on trouvera une variance vraie importante. Cependant, cette variance vraie n'est pas, nécessairement, une variance liée aux facteurs communs que l'on souhaite mesurer. Si on constate une augmentation de la fidélité, il ne s'ensuit pas, automatiquement, une augmentation de la validité. C'est, par exemple, le cas lorsqu'on administre des tests composés de QCM à deux populations distinctes, l'une composée d'élèves habitués à passer ce genre d'épreuves, l'autre ne l'étant pas (on enregistrera alors un effet lié à l'habitude de passer des QCM, en plus des facteurs que l'on souhaite mesurer). Les premiers risquent de répondre significativement mieux que les seconds à toutes les questions, même si leur compétence n'est pas réellement plus élevée[3].
- ***Motivation :*** certains sujets peuvent être plus ou moins motivés face au test. Par exemple, certains sont étudiants en psychologie et passent le test dans le cadre de leurs études, parce qu'ils sont inscrits à un " subjects pool ", mais sans que les résultats ne puissent modifier leur carrière, alors que d'autres résultats proviennent d'étudiants soumis à ce test dans le cadre d'un examen d'entrée ou d'un test d'embauche.
- ***Dispositions mentales :*** certains sujets peuvent, par exemple, être dans des conditions mentales particulières de fatigue ou de stress, ce qui risque d'altérer la mesure.

[3] Cet effet n'est cependant pas automatique et aussi simple que cela puisque, par exemple, les élèves belges francophones des deux premières années d'enseignement secondaire soumis en 1995 à une étude internationale en mathématiques et en sciences réussissent relativement mieux les QCM que les questions ouvertes, contrairement aux idées triviales sur le sujet, basées sur la plus grande familiarité de ces élèves avec le second type d'évaluations (Monseur et Demeuse, 1998).

- ***Choix au hasard (notamment dans les QCM, comme nous le verrons), impulsivité à répondre, rapidité (cf. vitesse/puissance) :*** certains sujets utilisent davantage le choix au hasard que d'autres lorsqu'ils ignorent la réponse.
- ***Disposition au travail :*** ce facteur est proche de la motivation, mais concerne le contenu ou la procédure de test, plutôt que le contexte. Ainsi, par exemple, certains sont à l'aise dans des situations en temps limité, alors que d'autres seront particulièrement stressés dans un tel cas.
- ***Effet de testing :*** la première passation d'un test permet d'améliorer le score à la seconde passation. Cette amélioration peut provenir d'une mémorisation des bonnes réponses ou de la méthode de résolution. On peut aussi observer, au contraire, une dégradation des résultats à cause d'une moins grande motivation à participer une nouvelle fois, ou une impulsivité plus grande (on ne fait pas très attention aux consignes, on oublie de se relire…).

Le choix du type de fidélité découle du type de variance d'erreur que l'on veut prendre en compte. Il est possible d'organiser différemment la liste des facteurs ci-dessus, de même qu'il est possible de les contrôler, au moins en partie, en tentant de définir de manière aussi précise que possible les situations de test et en tentant de les maintenir stables d'un sujet et d'une passation à l'autre.

2.3. Méthodes basées sur la consistance interne

Ces méthodes s'appliquent uniquement aux tests de puissance ou à ceux qui en approchent de très près les conditions, c'est-à-dire des instruments dont les différents items sont supposés mesurer parfaitement la même dimension et où l'on suppose que chaque sujet peut atteindre et résoudre l'ensemble des questions, selon son niveau de compétence, indépendamment du temps. Il est essentiel que tous les sujets aient le temps de répondre à tous les items du test, quelle que soit la position de ceux-ci dans le test. Toutes ces méthodes ne requièrent qu'une seule administration du test, ce qui en justifie l'emploi fréquent.

2.3.1. Méthodes " items pairs-impairs " (dite " odd-even " en langue anglaise)

Le procédé est le suivant : on calcule le score obtenu par chaque sujet à un test hypothétique qui serait composé des items pairs (n° 2-4-6...). On agit de même pour les items impairs (n°1-3-5…). On calcule ensuite la ***corrélation de Bravais-Pearson*** ρ_{pi} entre ces deux séries de scores. Cette corrélation constitue une bonne estimation de la fidélité d'un test de longueur moitié moindre que le test de départ, mais ayant la même variance vraie. Pour obtenir la fidélité du test de départ, ON DOIT CORRIGER la valeur du coefficient de corrélation de Bravais-Pearson en utilisant la ***formule de Spearman-Brown***, de manière à rétablir celle-ci en fonction de la longueur initiale du test. La fidélité du test initial ρ_{tt} sera donc estimée, à partir de ρ_{pi}, de la manière suivante:

$$\rho_{tt} = \frac{2\rho_{pi}}{1+\rho_{pi}}$$

On doit, cependant, être conscient qu'il y a ici parallélisme étroit (même moment d'administration, durée identique). Ce parallélisme est même trop fort. Il y a risque de surestimation car une variance vraie circonstancielle peut être prise pour de la variance vraie (ex. : tous les élèves distraits par un événement externe). C'est pourquoi certains préfèrent

utiliser des formes parallèles à un ou deux jours d'intervalle. Si le test se déroule en temps limité, l'effet parasite de la vitesse contribue à la variance vraie de façon égale pour les deux parties du test. Il y a donc surestimation de celle-ci, ce qui explique que la fidélité par consistance interne devrait être réservée aux tests de puissance.

2.3.2. La formule de Rulon

Rulon a développé une formule simple qui suit la définition de base de la fidélité (proportion de variance vraie dans un test).

$$\rho_{tt} = 1 - \frac{variance\ d'erreur}{variance\ totale}$$

Si les scores aux deux moitiés du test (items pairs et items impairs) sont obtenus pour chaque individu, on peut calculer pour l'ensemble des sujets la différence entre le score des items pairs et celui des items impairs de chacun d'eux. Ces différences représentent les erreurs. La variance calculée au départ de la distribution des erreurs σ^2_d représente la variance d'erreur σ^2_e.

On peut dès lors écrire que $\rho_{tt} = 1 - \frac{\sigma_d^2}{\sigma_t^2}$

Dans cette solution, on NE CORRIGE PAS par la ***formule de Spearman-Brown*** car on obtient la fidélité du score total et non des scores obtenus sur la moitié des items.

2.3.3. Les formules de Kuder-Richardson

Kuder et ***Richardson*** ont développé un groupe de procédures basées sur les statistiques d'items. En effet, on peut objecter aux deux méthodes " items pairs-impairs " décrites ci-dessus qu'il y existe un nombre très grand de manières de scinder un test de **k** items en deux parties de $\frac{k}{2}$ items et qu'il n'y a pas beaucoup de raisons de recourir à une procédure systématique particulière. On pourrait donc calculer un très grand nombre de ρ_{pi} différents, selon la méthode adoptée.

Kuder et Richardson proposent donc de découper le test en **k** parties de un item, ce qui résout le problème de l'arbitraire dans les partitions (comme c'est le cas dans la méthode pairs-impairs, par exemple). Le postulat de base est, dans ce cas, que tous les items sont factoriellement univoques et relèvent du même facteur commun.

Si $\mathbf{p_i}$ et $\mathbf{q_i}$ sont respectivement les proportions de réussite et d'échecs aux différents items, alors on peut écrire :

$$\rho_{tt} = \frac{k}{k-1}\left(\frac{\sigma_t^2 - \sum p_i q_i}{\sigma_t^2}\right)$$

Dans cette formule, baptisée ***KR20***, le numérateur de la fraction entre parenthèse représente la variance vraie, le dénominateur la variance totale et le rapport de k sur k-1 un coefficient de correction. Il est particulièrement utile de souligner que <u>*k est le nombre d'items dans le test*</u> et non le nombre de sujets auxquels le test est administré.

On peut considérer que le numérateur est une bonne estimation de la variance vraie, si on se rappelle que :

$$\sigma_t^2 = \sum p_i q_i + 2\sum \rho_{ij} \sqrt{p_i q_i p_j q_j}$$

Dès lors, on comprend que $\sigma_t^2 - \sum p_i q_i$ une bonne approximation de la variance vraie.

On utilise le terme de correction $\frac{k}{k-1}$ dès qu'il y a discrimination entre sujets et que, dès lors, $\Sigma p_i q_i \neq 0$. Cette correction est d'autant plus nécessaire que le nombre d'items est faible.

On obtient, en général, une fidélité plus basse avec le ***KR20*** qu'avec les méthodes pairs-impairs, car les conditions théoriques de base sont plus sévères (univocité des items) et rarement remplies.

On peut aussi utiliser le ***KR21***, moins précis, mais qui ne requiert pas le calcul de statistiques d'items. En effet, $\mathbf{p_m}$ – le pourcentage moyen de réussite - se dérive au départ de la moyenne des scores totaux sans qu'il ne faille calculer tous les $\mathbf{p_i q_i}$. La formule du ***KR21*** s'écrit :

$$\rho_{tt} = \frac{k}{k-1}\left(\frac{\sigma_t^2 - k p_m q_m}{\sigma_t^2}\right)$$

On peut noter que l'estimation de la fidélité par la méthode de ***KR21*** est inférieure ou égale à celle établie grâce au ***KR20***. En effet, $k p_m q_m \geq \Sigma p_i q_i$.

Les deux formules de Kuder et Richardson sont à présent supplantées par la formule de l'alpha de Cronbach, plus complexe à calculer, mais qui ne pose aucun problème aux ordinateurs actuels.

2.3.4. L'alpha de Cronbach

Cronbach a développé une formule plus générale que celle de Kuder et Richardson permettant de calculer la fidélité au départ de n'importe quel partionnement du test en **k** sous-tests, à condition que l'on assume l'univocité des **k** parties du test. On utilise aussi cette formule lorsqu'on traite des items notés de manière non dichotomiques, c'est-à-dire pouvant donner lieu à des notes qui ne se limitent pas à deux valeurs 0 et 1.

La ***formule de Cronbach*** constitue une généralisation du **KR20**. Elle s'écrit :

$$\rho_{tt} = \frac{k}{k-1}\left(\frac{\sigma_t^2 - \sum \sigma_k^2}{\sigma_t^2}\right)$$

Cette valeur est généralement notée α et appelée ***alpha de Cronbach***. Il est important de noter que k représente le nombre d'items (ou de groupes d'items, si l'analyse porte sur un regroupement d'items en sous-scores) et non le nombre de sujets.

2.3.5. L'approche de l'analyse de variance

Le résultat d'un sujet **s** à un item **i** peut être prédit par la difficulté spécifique de l'item **i** et par l'aptitude spécifique du sujet **s**.

Cependant, la prédiction n'est pas parfaite car il peut exister des erreurs dues à des interactions particulières entre les sujets et les items.

On peut donc identifier trois grandes sources de variance :

- $\mathbf{V_s}$: variance entre sujets.
- $\mathbf{V_i}$: variance entre items (n'intervient pas dans le raisonnement; les items sont, par essence, de difficulté variable).
- $\mathbf{V_{s*i}}$: variance liée à l'interaction items/sujets (cette source de variance est due au fait que les sujets ne répondent pas à chaque item comme cela est prédit par la difficulté de l'item et la compétence du sujet[4]).

Dans cette approche, on considérera que la variance d'erreur peut être assimilée à la part de variance liée à l'interaction entre les items et les sujets. La fidélité sera donc exprimée comme

$$\rho_{tt} = 1 - \frac{V_{s*i}}{V_s}$$

On notera que, dans ce cas, ρ_{tt} est égal à la fidélité mesurée par le ***KR20*** (puisqu'elle s'appuie sur les résultats aux items pris individuellement).

2.4. L'erreur standard de mesure

L'***erreur standard de mesure***, notée ***ESM***, permet de déterminer le degré de confiance que l'on peut accorder au score obtenu à un test donné par un sujet particulier. Elle est fonction de la qualité de l'instrument utilisé et donc de sa fidélité. Elle s'établit de la manière suivante.

$$ESM = \sigma_t \sqrt{1 - \rho_{tt}}$$

où $\boldsymbol{\sigma_t}$ est l'écart-type des résultats du test et $\boldsymbol{\rho_{tt}}$ la fidélité du test telle qu'elle a été calculée par l'une des méthodes abordées dans ce chapitre.

Exemple :

Un test de lecture a une moyenne de 48 points et un écart-type de 12 points. Un sujet X a obtenu le score de 34. La fidélité du test est de 0,64. Quelle est l'erreur standard de mesure ?

$$\mathbf{ESM = 12 \sqrt{1 - 0,64} = 12 \times 0,6 = 7,2}$$

Interprétation :

La fourchette (pour P≤ 0,05) sera de 1,96 x 7,2 points = 14,1 points par rapport au score de 34 points, soit un intervalle s'étendant de 19,9 points à 48,1 points. Si on est encore plus exigeant (P≤ 0,01), la fourchette sera de 2,58 x 7,2 points = 18,6 points par rapport au score de 34 points soit une fourchette plus large s'étendant de 15,4 points à 52,6 points. Il n'en demeure pas moins que le score de 34 points est la meilleure estimation de la compétence du sujet.

Le coefficient 1,96, dans la formule ci-dessus, pour P≤0,05, comme le coefficient 2,58, pour une valeur de P≤0,01, provient d'une table de distribution normale telle que celle fournie en annexe. On considère en effet que les erreurs de mesure se distribuent normalement. Il faut

[4] Dans certains cas, il est possible " d'expliquer " cette interaction par une ou plusieurs autres variables. Ainsi, le fait d'être plus ou moins familiarisé avec ce type de test peut-il conduire certains individus, à compétence égale, à mieux réussir certains items que d'autres, non familiarisés avec ce type d'examens. Dans d'autres cas, il est impossible de comprendre pourquoi une telle interaction apparaît.

garder à l'esprit le fait que, puisqu'il s'agit d'une fourchette, on doit tenir compte du caractère bilatéral de l'erreur.

Cette ***erreur standard de mesure*** est à rapprocher de l'***erreur d'échantillonnage*** qui est abordée dans les cours de statistique. Elle s'ajoute à celle-ci au sens où, généralement, on considère que les mesures ont été réalisées avec un instrument parfait et qu'on pratique une inférence, alors qu'à l'erreur liée à l'échantillonnage, on devrait encore adjoindre une erreur liée à la qualité de l'instrument. L'analyse de variance en mesures répétées, si on considère un même instrument administré plusieurs fois aux mêmes sujets, permet d'estimer à la fois l'importance de l'erreur de mesure et de l'erreur d'échantillonnage.

2.5. Utilité des coefficients de fidélité

Lorsqu'on mentionne les qualités psychométriques d'un test, c'est le coefficient de fidélité qui est le plus souvent cité. En effet, on établit beaucoup plus facilement un coefficient de fidélité qu'un coefficient de validité externe, comme nous allons le voir dans le chapitre suivant, relatif à la validité.

Cependant, même si la validité est primordiale, il est également important d'établir la fidélité. En effet, pour des tests diagnostiques, qu'ils soient psychologiques ou pédagogiques, une bonne fidélité est essentielle : à quoi bon sinon de mesurer quelque chose de bien défini si cette mesure est entachée d'une erreur importante ! De même, il est essentiel de s'assurer d'une bonne fidélité lorsqu'il s'agit de mesurer un changement dans le temps à l'aide d'un même instrument. Faute d'une excellente fidélité, le chercheur sera amené à attribuer au temps ou à un traitement intervenu entre les deux mesures un effet purement fortuit, lié à l'instabilité de son instrument de mesure.

Une augmentation de la fidélité correspond souvent à une augmentation de la validité, mais la relation n'est pas aussi simple qu'il y paraît. En effet, une augmentation de la validité implique, dans certains cas, l'introduction de mesures qui prennent en compte différents facteurs communs. Il y a donc, dans ce cas, diminution de la fidélité, du moins si on la mesure par des méthodes liées à la consistance interne. C'est pourtant de cette manière qu'on procède généralement, même si c'est inadéquat dans ce cas précis, car cette méthode n'implique qu'une seule administration d'une seule et même forme de test.

La fidélité du critère, lors d'expériences de validation, pose un autre problème important. En effet, si le critère est peu fidèle (ex. : résultats à des examens traditionnels, surtout si les questions sont ouvertes et conduisent à des réponses longues, comme dans le cas de dissertations), il est difficile de trouver des prédicteurs corrélant bien avec lui, même si ceux-ci ont une bonne fidélité.

2.6. Interprétation de la valeur des coefficients de fidélité

En dehors de l'utilisation du coefficient de fidélité dans le calcul de l'erreur standard de mesure, relativement aisée à interpréter en elle-même, les valeurs obtenues par les différentes formules qui ont été évoquées parlent peu aux personnes qui ne sont pas habituées à la construction de tests. Laurencelle (1998, pp. 93-94) propose plusieurs solutions pour donner du sens à ces indices La plus simple et la plus pragmatique des solutions proposées est sans doute le recours à un barème particulier et communément admis dans le domaine considéré (test de performance, échelles d'attitude…). Pour ce qui concerne les tests de performance, on peut adopter le barème suivant (tableau 2).

Tableau 2 – *Barème d'appréciation du coefficient de fidélité dans le cas des tests de performance (d'après Laurencelle, 1998, p. 94).*

Valeur de ρ_{tt}[5]	Appréciation
0,95 à 1,00	Instrument parfait, les mesures sont pratiquement sans erreur.
0,85 à 0,95	Instrument excellent, les mesures contiennent peu d'erreur.
0,70 à 0,85	Bon test, il est prudent d'évaluer une seconde fois le sujet.
0,50 à 0,70	Instrument imprécis, peut contenir de l'information utile.
0,00 à 0,50	Instrument peu utile, ne pas l'employer pour classer un sujet.

Le barème qui est proposé ci-dessus est certainement trop sévère pour les échelles d'attitude où les coefficients de fidélité observés sont assez généralement plus bas, mais trop peu sévère pour des mesures anthropométriques. Il s'agit donc de prendre en compte le domaine d'intérêt, même si la taille de l'erreur standard de mesure reste bien affectée par la valeur de la fidélité, quel que soit le domaine dans lequel on travaille. En procédant de la sorte, on admet donc simplement que la taille de l'erreur peut être plus ou moins importante, selon le type de mesure effectuée, faute de mieux. Il ne s'agit donc pas d'une solution parfaitement satisfaisante.

2.7. Problèmes spécifiques relatifs à la fidélité

Jusqu'à présent, nous avons principalement traité des différentes approches permettant d'estimer la fidélité des tests. Il convient également d'analyser un certain nombre de problèmes particuliers (ex. : relations entre fidélité et dispersion des aptitudes des sujets) ainsi que de préciser comment s'effectue le calcul de la fidélité pour certaines mesures particulières (ex. : mesures composites...).

2.7.1. Conditions optimales de difficulté

On a vu que l'on obtient une discrimination maximale entre les sujets à deux conditions : tous les $\mathbf{p_i = 0,5}$ et tous les $\boldsymbol{\rho_{ij} = 1}$. On a aussi expliqué que ces conditions n'étaient jamais totalement remplies dans la pratique. Néanmoins, si on veut discriminer au mieux les élèves supérieurs de ceux qui sont inférieurs à la moyenne, c'est ce genre de condition qu'il faut rechercher ($\mathbf{p_i}$ voisins de 0,50 et inter-corrélations élevées entre items). Il peut arriver aussi que l'on veuille discriminer en un point précis de la distribution (ex. : concours où on l'on retient les 25 % des sujets qui présentent les meilleurs résultats). Dans ce cas, les $\mathbf{p_i}$ doivent être ciblés pour que le sujet situé au point de coupure ait une probabilité de 0,50 de répondre à chaque item. Ces items doivent également présenter des inter-corrélations élevées.

Par ailleurs, on désire très souvent obtenir à la fois une bonne discrimination entre sujets (sur toute la distribution des aptitudes) ainsi qu'une fidélité élevée. C'est ce qu'on obtiendra si on choisit des items de difficulté variable présentant des inter-corrélations moyennes.

(Remarque : les inter-corrélations maximales entre items de difficultés variables ne valent pas 1, mais décroissent très vite[6] en cas de variation importante des $\mathbf{p_i}$).

5 Comme il s'agit d'un ordre de grandeur, l'auteur ne s'est pas préoccupé de fournir des classes dont les bornes sont clairement identifiées.

6 Ceci est dû au fait que le coefficient de corrélation entre deux variables dichotomiques (coefficient □) n'a pas nécessairement 1 pour borne supérieure.

2.7.2. Fidélité des tests de vitesse

Les méthodes de calcul de la fidélité fondées sur la consistance interne ne s'appliquent qu'aux tests de puissance, nous l'avons déjà mentionné. On a donc dû développer une procédure spéciale permettant d'évaluer la fidélité pour des tests de vitesse.

C'est ainsi qu'on peut appliquer les deux moitiés du test en temps limité – la moitié du temps total - et à deux moments proches. On corrigera ensuite par Spearman-Brown. Mais, d'une manière générale, ce sont les méthodes basées sur le test-retest ou les formes parallèles qui sont utilisées, impliquant plusieurs passations si la formule consistant à tester séparément deux moitiés du test original ne peut être retenue.

2.7.3. Fidélité et dispersion des aptitudes dans la population

Lorsque la fidélité d'un test a été établie sur une première population dont la variance est égale σ_1^2 et que l'on utilise ce même test sur une deuxième population présentant des caractéristiques différentes (σ_2^2), on peut s'attendre à observer des variations de la fidélité, liées à la différence de répartition des aptitudes dans les deux populations.

On pourra estimer la fidélité d'un test présenté à une population donnée, soit ρ_{22}, à partir de celle qui a pu être établie sur une autre population, soit ρ_{11}, si on connaît la variabilité de la compétence mesurée par le test dans les deux populations (soit σ_1^2 et σ_2^2), grâce à la formule ci-dessous.

$$\rho_{22} = 1 - \frac{\sigma_1^2 (1 - \rho_{11})}{\sigma_2^2}$$

Si la seconde population présente une moins grande dispersion que celle de la première population, sur laquelle la fidélité a pu être calculée, on observera une diminution de la fidélité de l'instrument, lorsqu'il sera utilisé avec cette seconde population. Au contraire, si la compétence est plus largement dispersée dans la seconde population, c'est à un accroissement de la fidélité de l'instrument que l'on devra s'attendre.

Exemple

Au départ des trois paramètres suivants, connus

$$\sigma_1^2 = 2 \qquad \rho_{11} = 0,8 \qquad \sigma_2^2 = 1$$

On pourra estimer la fidélité du test lorsqu'il sera administré à une seconde population dont la dispersion de la compétence est connue par ailleurs. Dans l'exemple, cette fidélité sera estimée de la manière suivante :

$$\rho_{22} = 1 - \frac{2}{1}(1 - 0,8) = 1 - 2\,(0,2) = 0,6$$

2.7.4. Fidélité des tests à choix multiple (QCM)

Les QCM classiques sont moins fidèles que les tests à réponses ouvertes, du fait que l'élément chance est susceptible d'intervenir. Ceci est d'autant plus vrai que le nombre de distracteurs est petit. Si on augmente le nombre de distracteurs, la fidélité augmentera puisque le " facteur

chance " est moins important, à la condition que les distracteurs ajoutés aient même valeur discriminative que les originaux. En pratique, lorsqu'on utilise des questions à choix multiple, les $\mathbf{p_i}$ seront en moyenne supérieurs à 0,50. Nous traiterons ce problème dans le chapitre 5, consacré à l'analyse d'item.

2.7.5. Fidélité des scores composites

Cette mesure de la fidélité est importante, lorsqu'on utilise le score global obtenu à une batterie de tests, comprenant des sous-tests mesurant des aptitudes différentes. Lorsque ces sous-tests sont utilisés pour former un score composite, la fidélité de chacun des sous-tests peut être plus faible que ce que l'on exigerait de chaque sous-test s'il était utilisé isolément en vue d'un usage diagnostique.

On doit à ***Mosier*** (Guilford, 1954) la formule qui permet d'estimer la fidélité d'un score composite ρ_{cc} à partir des fidélités de chacun des sous-tests.

$$\rho_{cc} = 1 - \frac{\sum p_j^2 \sigma_j^2 - \sum p_j^2 \sigma_j^2 \rho_{jj}}{\sum p_j^2 \sigma_j^2 + 2 \sum p_j p_k \sigma_j \sigma_k \rho_{jk}}$$

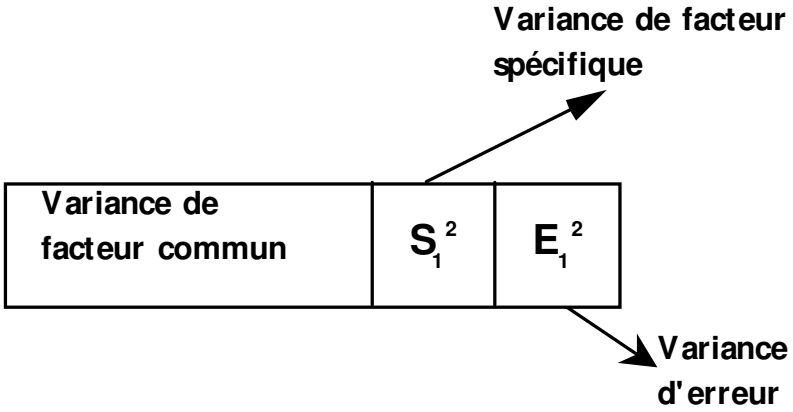
Cette formule permet de calculer la fidélité du score composite obtenu en combinant deux ou plusieurs tests. Les indices **j** et **k** font référence à deux tests particuliers, la valeur de k étant supérieure à j. L'application de cette formule nécessite la connaissance des éléments suivants pour chacun des tests : la variance du score total (σ^2_j), la fidélité (ρ_{jj}) et le poids, élevé au carré, accordé à chaque test (p^2_j). Il convient encore de disposer de toutes les inter-corrélations (ρ_{jk}) entre les scores totaux de chacun des tests qui composent l'ensemble.

2.7.6. Fidélité des scores différentiels

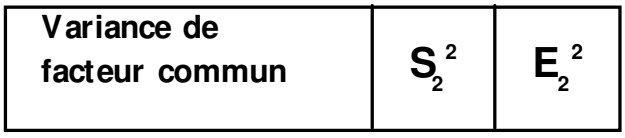
Lorsqu'on est confronté à des problèmes de guidance scolaire ou d'orientation professionnelle, le problème se présente souvent en termes de profils et/ou de scores différentiels. A ce moment, se pose le problème de la fidélité des scores différentiels. On peut établir que cette fidélité des scores différentiels est, de loin, plus basse que les fidélités des scores au départ desquels sont calculées les différences. On peut aussi montrer que, plus la fidélité de chacun des scores de départ est élevée, plus haute sera la fidélité des scores différentiels. On comprendra aussi aisément qu'une corrélation élevée entre les scores de départ conduit à un abaissement de la fidélité des scores différentiels.

On peut représenter ce phénomène comme suit :

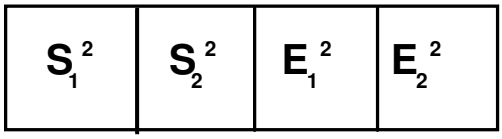
Représentation des variances dans le test 1.



Représentation des variances dans le test 2.



Représentation des variances du score $X_1 - X_2$



La part de variance vraie, soit $S_1^2 + S_2^2$, est donc faible par rapport aux variances d'erreur, une fois la variance commune enlevée.

On obtient la fidélité du score différentiel pour les tests j et k en calculant :

$$\rho_{dd} = \frac{\rho_{jj} + \rho_{kk} - 2\rho_{jk}}{2(1 - \rho_{jk})}$$

où ρ_{dd} est la fidélité du score différentiel $X_j - X_k$.

ρ_{jj} et ρ_{kk} sont, respectivement, la fidélité des tests j et k

ρ_{jk} est la corrélation entre les scores totaux des tests j et k, soit X_j et X_k.

Exemple

Soit deux tests (j et k). On souhaite connaître la valeur qui sépare les deux scores. La fidélité du premier test est de 0,8, alors que celle du second est de 0,7.

$\rho_{jj} = \mathbf{0,8}$

$\rho_{kk} = \mathbf{0,7}$

La corrélation entre les deux scores des tests j et k s'élève à 0,3.

$\rho_{jk} = \mathbf{0,3}$

La fidélité du score calculé à partir de la différence entre le score au test j et le score au test k s'élève à 0,64 et se calcule de la manière suivante:

$$\rho_{dd} = \frac{0,8 + 0,7 - 0,6}{2 \,.\, (1 - 0,3)} = \frac{0,9}{1,4} = 0,64$$

On peut encore observer que si $\rho_{jk} = \frac{\rho_{jj} + \rho_{kk}}{2}$, c'est-à-dire si la corrélation des deux tests est égale à leur fidélité moyenne, alors $\rho_{dd} = 0$. Par contre, si $\rho_{jk} = 0$, c'est-à-dire si les deux tests sont non corrélés, alors $\rho_{dd} = \frac{\rho_{jj} + \rho_{kk}}{2}$. Dans ce cas dernier cas, la fidélité du score différentiel est égal à la moyenne des fidélités des deux tests de départ.

Bibliographie

Guilford, J.P. (1954). Psychometric Methods. New York : McGraw-Hill.

Laurencelle, L. (1998). *Théorie et techniques de la mesure instrumentale.* Sainte-Foy : Presses de l'Université du Québec.

Monseur, C., Demeuse, M. (1998). Apports des études internationales à la réflexion sur la qualité d'enseignement nationaux: une analyse de l'éducation scientifique en Communauté française de Belgique. *Bulletin de la Société Royale des Sciences de Liège, 67(5),* 261-280.