Mesure et évaluation en éducation

Méthodes d'analyse d'items et optimisation de la fiabilité des mesures en éducation

Gratien Bambanota Mokonzi

Volume 26, Number 1-2, 2003

Généralisabilité

URI: https://id.erudit.org/iderudit/1088240ar
DOI: https://doi.org/10.7202/1088240ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

Bambanota Mokonzi, G. (2003). Méthodes d'analyse d'items et optimisation de la fiabilité des mesures en éducation. *Mesure et évaluation en éducation*, *26*(1-2), 61–73. https://doi.org/10.7202/1088240ar

Article abstract

This study compares the contribution of item analysis methods developed in classical psychometrics and in the context of generalizability theory for the optimization of the dependability of educational measures. To this end, it shows which method to use in order to increase reliability or to reduce the impact of random fluctuations affecting the measure. Apart from the conclusions reached, the study raises questions that could be addressed through further research.

Méthodes d'analyse d'items et optimisation de la fiabilité des mesures en éducation

Gratien Bambanota Mokonzi
Faculté de psychologie et des sciences de l'éducation, Université de Kisangani

MOTS CLÉS : Fiabilité, fidélité, généralisabilité, optimisation, psychométrie, évaluation scolaire, analyse d'items, analyse de facettes, covariance, analyse de la variance

L'étude confronte les méthodes d'analyse d'items mises au point en psychométrie classique et dans le modèle de la généralisabilité, du point de vue de l'optimisation de la fiabilité des mesures scolaires. Elle indique à cet effet quelle méthode utiliser si l'on veut accroître la fidélité ou encore réduire l'ampleur des fluctuations aléatoires affectant la mesure. Au-delà des résultats auxquels elle aboutit, l'étude soulève des questions que l'on pourrait examiner dans des recherches ultérieures.

KEY WORDS : Dependability, reliability, generalizability, optimization, psychometrics, school assessment, item analysis, facet analysis, covariance, analysis of variance

This study compares the contribution of item analysis methods developed in classical psychometrics and in the context of generalizability theory for the optimization of the dependability of educational measures. To this end, it shows which method to use in order to increase reliability or to reduce the impact of random fluctuations affecting the measure. Apart from the conclusions reached, the study raises questions that could be addressed through further research.

PALAVRAS CHAVE: Fiabilidade, fidelidade, generalizabilidade, optimização, psicometria, avaliação escolar, análise de itens, análise de facetas, covariância, análise da variância

Este estudo confronta os métodos de análise de itens desenvolvidos na psicometria clássica e no modelo da generalizabilidade, do ponto de vista da optimização da fiabilidade das medidas escolares. Para esse efeito ele indica o método a utilizar se se pretender aumentar a fidelidade ou reduzir a amplitude das flutuações aleatórias que afectam a medida. Para além dos resultados a que chega, o estudo levanta questões que poderão ser examinadas em investigações ulteriores.

Problématique de l'analyse d'items

La préoccupation centrale de la psychométrie classique est la mise en évidence, de façon aussi précise que possible, des différences entre personnes. L'une des procédures préconisées pour la maximisation de ces différences consiste à éliminer, au cours du processus d'élaboration des instruments de mesure, des questions dont le pouvoir discriminatif, mesuré par la covariance item-test, est faible.

Cependant, l'accroissement des différences intersujets ne concorde pas toujours avec le but principal d'une épreuve pédagogique qui n'est pas nécessairement le classement des élèves, mais qui peut être aussi la différenciation des objectifs, la détermination du progrès d'apprentissage, etc. Ainsi, la procédure classique d'analyse d'items conduit à biaiser l'objet de l'évaluation en éducation, d'autant plus qu'elle privilégie les questions qui mesurent l'intelligence au détriment des questions qui mesurent simplement l'apprentissage (parce que celles-ci, en fin d'études, sont réussies généralement par tous les élèves) (Cardinet, 1986, pp. 114-115).

L'inadéquation de la méthode de corrélation aux différentes finalités de l'évaluation pédagogique a amené certains psychométriciens, tels que Cardinet, Tourneur, Bertrand et d'autres, à exploiter, dans le cadre de la théorie de la généralisabilité, une autre procédure d'analyse d'items, qu'ils appellent l'analyse de facettes. Visant la réduction de l'erreur, cette méthode fait ressortir la part de chaque niveau de la face d'instrumentation (des conditions d'observation) et de la face de différenciation (des objets d'étude) à la variance d'erreur et rejette les niveaux dont la contribution est très élevée.

Une troisième méthode d'analyse d'items, élaborée par Bertrand, sélectionne les niveaux d'une facette en considérant leur effet total sur la fidélité d'un instrument de mesure, en s'appuyant, pour cela, sur leur apport à la fois à la variance d'erreur et à la variance vraie.

Quel effet ces trois directions d'analyse d'items produisent-elles sur la fiabilité des mesures lorsqu'elles sont appliquées, après l'étude de la fidélité sur l'échantillon observé, pour la réduction de l'univers de généralisation ? Telle est la question à laquelle s'intéresse cette étude.

Hypothèses de la recherche

L'amélioration de la fidélité de la mesure nécessite l'accroissement de la part de la variance vraie dans la variance totale, lequel accroissement provient soit de l'augmentation de l'importance de la variance vraie, soit de la diminution de la contribution de la variance d'erreur. En sélectionnant les items les plus corrélés avec le score total[1], la méthode de covariance classique en psychométrie, symbolisée dans cette présentation par Max Covar, vise la maximisation de la variance vraie.

L'analyse de facettes, au contraire, se préoccupe de la réduction de la variance d'erreur, en sélectionnant les items les plus homogènes à l'ensemble. Sa démarche consiste à calculer la part de la variance d'interaction des faces de différenciation (face D) et d'instrumentation (face I) attribuable à chaque niveau de la face d'instrumentation[2], c'est-à-dire la contribution de chaque niveau à l'erreur relative. En ajoutant à cette contribution la part de la variance entre les niveaux de la face I qui revient au niveau considéré du fait de sa moyenne, on obtient la part de la variance d'erreur absolue due à chaque niveau. La diminution de chaque variance d'erreur (relative et absolue) découle de l'élimination des niveaux qui y contribuent le plus. Ces deux perspectives de l'analyse de facettes sont représentées dans ce texte par les codes Min $\sigma(\delta)$ et Min $\sigma(\Delta)$.

Contrairement aux méthodes précédentes, la maximisation de la fidélité relative (Max $\rho^2(\delta)$) ou absolue (Max $\rho^2(\Delta)$) analyse l'effet total de chaque niveau d'une facette sur la fidélité de l'instrument de mesure en considérant son apport à la fois à la variance vraie et à la variance d'erreur. L'analyse indique alors par quel niveau il faut amorcer le rejet des éléments d'une facette pour optimiser la fidélité de la mesure.

Compte tenu des fondements théoriques de ces différentes méthodes, cette étude a été orientée par les hypothèses suivantes :

- Par rapport à la fidélité initiale, chaque méthode d'optimisation améliorera, mieux que les autres méthodes, la caractéristique de la mesure qu'elle vise spécifiquement. Autrement dit,
  - la méthode qui vise à augmenter la variance des scores vrais (Max Covar) sera celle qui y parviendra le mieux ;
  - les méthodes qui visent la diminution de l'erreur (Min $\sigma(\delta)$ et Min $\sigma(\Delta)$) seront celles qui y parviendront le mieux, séparément pour chaque type d'erreur ;

- les méthodes qui visent à maximiser la fidélité de l'épreuve optimisée (Max $\rho^2(\delta)$, Max $\rho^2(\Delta)$) seront celles qui y parviendront le mieux, séparément pour chaque type de fidélité.

- Les méthodes d'optimisation n'amélioreront pas nécessairement les éléments caractéristiques de la mesure sur les points qu'elles ne visent pas spécifiquement.

Démarche méthodologique

Le plan d'observation

Les données exploitées se rapportent aux résultats obtenus par 184 élèves à une épreuve de mathématiques de 64 questions appliquée avant et après l'enseignement des notions ciblées[3]; elles sont structurées suivant un plan d'observation à trois facettes croisées (Sujets, Items et Phases).

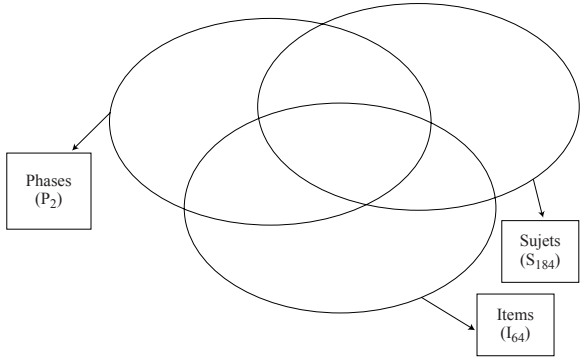

Figure 1. ***Représentation du plan d'observation S_{184} x P_2 x I_{64}***

Analyse des données

Bien que les méthodes d'analyse d'items soient applicables symétriquement sur la face de différenciation et sur la face d'instrumentation, leur contribution a été étudiée uniquement à partir de la réduction de l'univers de généralisation, pour ne faire varier qu'un facteur à la fois. Pour cela, les caractéristiques de la fidélité des données initiales ont été comparées à celles des données sélectionnées par chaque méthode. Cinq caractéristiques ont été prises en compte pour cette comparaison : l'écart type des scores univers ($\sigma(\tau)$), l'écart type de l'erreur relative ($\sigma(\delta)$), l'écart type de l'erreur absolue ($\sigma(\Delta)$), le coefficient de généralisabilité relatif ($\rho^2(\delta)$) et le coefficient de généralisabilité absolu ($\rho^2(\Delta)$).

Pour assurer une comparaison équitable des méthodes, nous avons éliminé pour chacune d'elles les 10% des niveaux les moins bons de la face de généralisation. Du fait de cette réduction uniforme de la taille des échantillons observés, il n'a pas été nécessaire d'appliquer la correction de Spearman-Brown pour comparer les fidélités résultantes. Mais pour comparer les résultats de chaque méthode aux valeurs initiales, on devra se souvenir que l'effectif des échantillons observés (par exemple, le nombre de questions du test) a été réduit de 10%.

Toutes ces analyses ont été effectuées pour trois finalités différentes de la mesure en éducation : la discrimination des sujets, la différenciation des questions et la comparaison des phases d'apprentissage.

Présentation des résultats

Méthodes d'analyse d'items et différenciation des sujets

Les résultats obtenus après l'appréciation de la fidélité de l'épreuve de 64 questions, le rejet de six items et l'estimation de la fidélité de l'épreuve optimisée par chaque méthode sont repris dans les tableaux 1 et 2.

Tableau 1
Fidélité pour la différenciation des sujets au prétest

Coefficients	*Valeurs initiales*	*Min σ (δ)*	*Min σ (Δ)*	*Max Covar*	*Max $\rho^2(\delta)$*	*Max $\rho^2(\Delta)$*
σ (τ)	0,106664	0,101098	0,101098	0,118575	0,117620	0,113825
σ (δ)	0,040274	0,039774	0,039774	0,043509	0,042228	0,042410
σ (Δ)	0,043111	0,041432	0,041432	0,046379	0,045426	0,045042
ρ^2 (δ)	0,875225	0,865967	0,865967	0,881339	0,885821	0,878102
ρ^2 (Δ)	0,859579	0,856199	0,856199	0,867314	0,870203	0,864610

Tableau 2
Fidélité pour la différenciation des sujets au posttest

Coefficients	*Valeurs initiales*	*Min σ (δ)*	*Min σ (Δ)*	*Max Covar*	*Max $\rho^2(\delta)$*	*Max $\rho^2(\Delta)$*
σ (τ)	0,157454	0,163376	0,166638	0,170207	0,168197	0,170207
σ (δ)	0,052049	0,053669	0,054649	0,055006	0,054272	0,055006
σ (Δ)	0,059364	0,062141	0,061874	0,061863	0,061971	0,061863
ρ^2 (δ)	0,901489	0,902597	0,902894	0,905436	0,905701	0,905436
ρ^2 (Δ)	0,875544	0,873614	0,878836	0,883311	0,880477	0,883311

Les caractéristiques initiales de la fidélité attestent que l'apprentissage réalisé entre le prétest et le posttest a eu comme effet l'accentuation des différences entre les élèves, l'écart type des scores univers ayant considérablement augmenté au second moment de la mesure par rapport au premier moment (soit un accroissement de 48%). Ce renforcement des différences inter sujets se reflète également à travers l'étendue de variation des moyennes des sujets qui est respectivement de 0,59375 au prétest et de 0,828125 au posttest (*cf.* figure 2). Comme le note Bain (à paraître dans ce numéro), bien qu'elle soit à l'encontre de la pédagogie de maîtrise, cette tendance, très couramment observée, tient à la difficulté de gérer pédagogiquement les différences d'origines diverses entre les élèves.

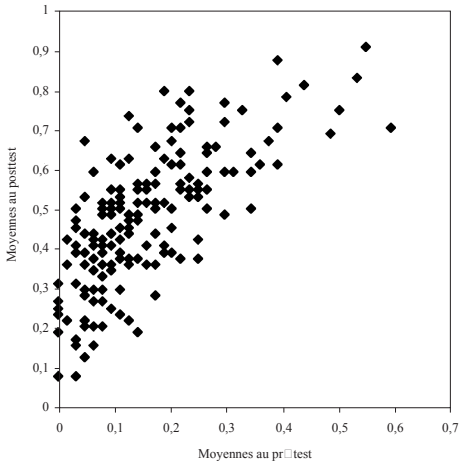

Figure 2. ***Distribution des moyennes des sujets aux deux phases de la mesure***

Parallèlement, la variance d'interaction Sujets x Questions (SQ), déterminant l'écart type de l'erreur relative $\sigma(\delta)$, et la variance interquestions, qui s'ajoute à la variance d'interaction SQ pour donner la variance d'erreur absolue, conduisant à $\sigma(\Delta)$, ont également augmenté au posttest, mais moins fortement que l'écart type des scores univers, entraînant ainsi l'augmentation des deux coefficients de fidélité.

Concernant l'effet des méthodes d'analyse d'items sur la fiabilité, en éliminant le même nombre d'items (soit six), les méthodes de covariance (Max Covar) et de maximisation des coefficients de fidélité (Max $\rho^2(\delta)$ et Max $\rho^2(\Delta)$) ont augmenté, au prétest (tableau 1), aussi bien les valeurs de l'écart type

des scores univers, des coefficients de généralisabilité que celles des écarts types des erreurs relative et absolue. L'amélioration la plus importante de $\sigma(\tau)$ a été enregistrée par Max Covar, alors que l'augmentation la plus sensible des coefficients de généralisabilité relatif et absolu a été obtenue par Max $\rho^2(\delta)$. De leur côté, Min $\sigma(\delta)$ et Min $\sigma(\Delta)$ sont les seules à avoir réduit les écarts types des erreurs relative et absolue, mais elles ont en même temps diminué, plus fortement encore, l'écart type des scores univers, provoquant ainsi la baisse des coefficients de fidélité. Ceci est dû au fait que ces méthodes ont éliminé à la fois des questions différenciatrices et non différenciatrices des sujets.

Au posttest (tableau 2), Max Covar et Max ρ^2 (Δ) ont le mieux augmenté les valeurs de l'écart type des scores univers et du coefficient de généralisabilité absolu, tandis que Max ρ^2 (δ) a le mieux optimisé le coefficient de fidélité relatif. Mais comme au prétest, la sélection des items par ces trois méthodes a augmenté les écarts types des erreurs. Ces dernières caractéristiques ont également été amplifiées par Min $\sigma(\delta)$ et Min $\sigma(\Delta)$, mais dans une moindre mesure. En effet, Min σ (δ) a bien minimisé, comme attendu, la valeur de σ (δ), et Min σ (Δ) celle de σ (Δ), à quelques décimales près.

Méthodes d'analyse d'items et différenciation des questions

Améliorer la différenciation des questions par la réduction de l'univers de généralisation implique, selon le plan de mesure I_{64}/S_{184}, l'application des méthodes d'analyse d'items à la facette Sujets, laquelle se situe alors sur la face d'instrumentation. Il s'agit ainsi d'écarter les sujets qui ne permettent pas de bien différencier les questions.

Pour cette finalité de la mesure, les caractéristiques de la fidélité de base et celles de la fidélité des mesures optimisées par chaque méthode, après le rejet de 18 sujets, sont reprises dans les tableaux 3 et 4.

Tableau 3
Fidélité pour la différenciation des questions au prétest

Coefficients	*Valeurs initiales*	*Min σ (δ)*	*Min σ (Δ)*	*Max Covar*	*Max ρ²(δ)*	*Max ρ²(Δ)*
σ (τ)	0,123057	0,116373	0,115946	0,135969	0,133560	0,132438
σ (δ)	0,023752	0,023632	0,023668	0,025765	0,024974	0,024967
σ (Δ)	0,025020	0,024199	0,024255	0,026989	0,026350	0,026085
ρ^2 (δ)	0,964083	0,960396	0,959996	0,965337	0,966216	0,965679
ρ^2 (Δ)	0,960302	0,958552	0,958074	0,962095	0,962534	0,962656

Tableau 4
Fidélité pour la différenciation des questions au posttest

Coefficients	*Valeurs initiales*	*Min σ (δ)*	*Min σ (Δ)*	*Max Covar*	*Max $\rho^2(\delta)$*	*Max $\rho^2(\Delta)$*
σ (τ)	0,228377	0,239355	0,240708	0,243822	0,242823	0,243805
σ (δ)	0,030697	0,031655	0,032404	0,032212	0,031892	0,032191
σ (Δ)	0,032818	0,034113	0,034044	0,033951	0,033996	0,033950
ρ^2 (δ)	0,982254	0,982811	0,982200	0,982846	0,983043	0,982865
ρ^2 (Δ)	0,979767	0,980093	0,980389	0,980980	0,980776	0,980978

L'amplification des différences entre les questions, occasionnée par l'apprentissage, se précise davantage lorsque cette facette est considérée comme objet d'étude. Pour un tel plan de mesure, l'écart type des scores univers a considérablement augmenté (de 0,123 à 0,228, soit 86% d'accroissement). Cela s'observe aussi à travers l'étendue de variation des moyennes des questions qui est passée de 0,505435 au prétest à 0,858696 au posttest (*cf.* figure 3). Les écarts types des erreurs de mesure ont également connu un accroissement, mais moins important que celui qu'a subi σ (τ), d'où l'amélioration des coefficients de généralisabilité.

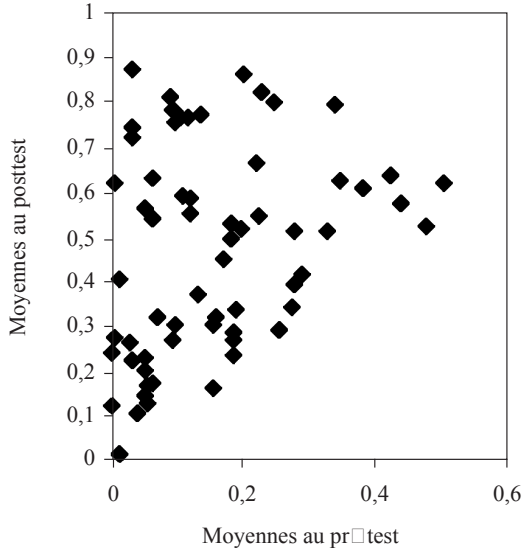

Figure 3. ***Distribution des moyennes des questions aux deux phases de la mesure***

Quant à la sélection des sujets, les méthodes de Max Covar, Max $\rho^2(\delta)$ et Max $\rho^2(\Delta)$ ont entraîné l'accroissement des valeurs de l'écart type des scores univers et des coefficients de généralisabilité (tableau 3). Encore une fois, l'augmentation la plus remarquable de l'écart type des scores univers a été accomplie par Max Covar. En outre, Max $\rho^2(\delta)$ et Max $\rho^2(\Delta)$ ont le mieux augmenté les coefficients de fidélité, la première ayant mieux optimisé la valeur de $\rho^2(\delta)$ et la seconde celle de $\rho^2(\Delta)$. Seules Min $\sigma(\delta)$ et Min $\sigma(\Delta)$ ont occasionné la diminution des valeurs des écarts types des erreurs relative et absolue. Mais, cette diminution s'est accompagnée d'une réduction, plus forte, des valeurs de $\sigma(\tau)$, entraînant la diminution des coefficients de fidélité (mais pour un échantillon de sujets de 10 % plus faible).

La seconde phase d'apprentissage (tableau 4), a suscité l'accroissement de toutes les caractéristiques de la mesure considérées dans cette étude, à une exception près (Min $\sigma(\Delta)$ n'a pas augmenté la valeur de ρ^2 (δ)). La valeur de l'écart type des scores univers est la plus élevée pour les données sélectionnées par Max Covar, même si elle ne diffère pas considérablement de l'écart type des scores univers des données sélectionnées par Max $\rho^2(\Delta)$. Max $\rho^2(\Delta)$ a optimisé, à égalité avec Max Covar, le coefficient de généralisabilité absolu, tandis que Max $\rho^2(\delta)$ a le mieux amélioré le coefficient de généralisabilité relatif.

Méthodes d'analyse d'items et différenciation des phases d'apprentissage

Pour cette troisième intention de la mesure, les analyses ont porté sur l'ensemble des données organisées selon le plan d'observation S_{184} X P_2 X I_{64} (*cf.* figure 1). Différencier les phases d'apprentissage sur la base de ce plan multifacettes revient à adopter le plan de mesure $P_2/S_{184}I_{64}$, c'est-à-dire le dispositif pour lequel les phases d'apprentissage sont des objets d'étude, les sujets et les items devenant les instruments de mesure.

En éliminant six items et 18 sujets par chaque méthode, nous avons obtenu les résultats ci-après.

Tableau 5
Fidélité pour la différenciation des phases

Coefficients	*Valeurs initiales*	*Min σ (δ)*	*Min σ (Δ)*	*Max Covar*	*Max ρ²(δ)*	*Max ρ²(Δ)*
σ (τ)	0,226198	0,198858	0,208108	0,263960	0,258004	0,254727
σ (δ)	0,020833	0,019011	0,020259	0,021504	0,021345	0,022605
σ (Δ)	0,025216	0,025896	0,025005	0,026414	0,026316	0,025462
ρ^2 (δ)	0,991589	0,990944	0,990612	0,993407	0,993202	0,992186
ρ^2 (Δ)	0,987726	0,983325	0,985769	0,990085	0,989703	0,990107

En rapport avec les hypothèses de cette recherche, les résultats repris dans le tableau 5 sont assez nets : toutes les méthodes d'optimisation ont amélioré les éléments caractéristiques de la mesure sur les points qu'elles visent spécifiquement. De plus, elles n'ont pas nécessairement amélioré les paramètres de la mesure qu'elles ne visent pas spécifiquement. Ceci est vrai pour Min $\sigma(\delta)$ et Min $\sigma(\Delta)$ qui ont réduit les valeurs de l'écart type des scores univers et celles des coefficients de fidélité par rapport aux valeurs initiales. C'est également vrai pour Max Covar, Max $\rho^2(\delta)$ et Max $\rho^2(\Delta)$ qui, au lieu de diminuer, ont plutôt augmenté les valeurs des écarts types des erreurs.

Pourtant, contrairement à ce que nous avions supposé au départ de cette étude, chaque méthode d'optimisation n'a pas toujours amélioré mieux que les autres méthodes la caractéristique de la mesure qu'elle visait particulièrement. En effet, pour le coefficient de généralisabilité relatif, c'est Max Covar, et non Max $\rho^2(\delta)$ qui en a le mieux augmenté la valeur. La différence est cependant infime.

Synthèse des résultats

Le tableau 6 permet de synthétiser les résultats des cinq tableaux précédents. Il indique les méthodes trouvées optimales (colonnes) pour chaque intention (en ligne). Dans les cases sont notés les numéros des études menant à ces résultats.

Tableau 6
Évaluation des méthodes d'analyse d'items

Intentions	*Min σ (δ)*	*Min σ (Δ)*	*Max Covar*	*Max $\rho^2(\delta)$*	*Max $\rho^2(\Delta)$*
Maximiser $\sigma(\tau)$			1), 2), 3), 4), 5).	2).	
Minimiser $\sigma(\delta)$	1), 2), 3), 4), 5).	1).			
Minimiser $\sigma(\Delta)$	1), 3).	1), 5).	2).		2), 4).
Maximiser $\rho^2(\delta)$			5).	1), 2), 3), 4).	
Maximiser $\rho^2(\Delta)$			2), 4).	1).	2), 3), 4), 5).

Ce tableau montre quelle méthode s'est révélée la meilleure pour chaque intention, dans les cinq situations que nous avons traitées. Il permet ainsi de déterminer expérimentalement quelle méthode utiliser si l'on veut, par exemple, maximiser la variance des scores vrais. Dans les cinq études précédentes, c'est la méthode d'analyse d'items classique, s'appuyant sur la covariance entre l'item et la note totale, qui a atteint ce résultat supérieur, et cela quelle que soit la visée de la mesure (différencier des sujets, des items, ou des phases).

Si l'on veut minimiser l'erreur relative, les cinq études montrent tout aussi clairement qu'il faut utiliser la méthode d'analyse d'items qui élimine les niveaux causant la plus grande variance d'interaction entre les faces de différenciation et d'instrumentation.

Les résultats sont moins nets lorsqu'on veut minimiser la variance d'erreur absolue. La palme n'est revenue à la méthode prévue dans ce but, Min σ (Δ), que dans les tableaux 1 et 5. Des résultats équivalents semblent avoir été obtenus deux fois par la méthode qui maximise ρ^2 (Δ), et d'autres fois par d'autres méthodes encore. Il est possible que l'importance relative de la variance d'interaction et de la variance spécifique (entre les niveaux de la facette d'instrumentation) joue un rôle dans ces variations.

Pour maximiser la fidélité relative (ρ^2 (δ)), la méthode à utiliser ne fait guère de doute, à considérer nos résultats. Dans quatre cas sur cinq, c'est la méthode prévue pour cela qui donne bien la fidélité la plus élevée.

Enfin, pour maximiser la fidélité absolue, c'est aussi dans quatre cas sur cinq la méthode prévue à cet effet qui donne les meilleurs résultats (si l'on admet les *ex æquo* comme cas favorables).

Conclusions

Dans la phase d'optimisation de la mesure, le coefficient de fidélité, qui n'est, à la limite, qu'un rapport abstrait, ne fournit pas de renseignements aussi précieux que ceux qu'apportent les paramètres de fidélité que sont l'écart type des scores univers, l'écart type de l'erreur relative et l'écart type de l'erreur absolue. L'objectif de cette étude était de déterminer quelle méthode d'analyse d'items utiliser pour renforcer chacune de ces caractéristiques de la mesure (le coefficient de fidélité y compris).

Au terme des analyses effectuées, il s'avère que, puisqu'elle maximise le mieux la variance des objets d'étude, la méthode de covariance item-test, élaborée en psychométrie classique, convient le mieux pour la construction

des épreuves normatives. De même, étant donné qu'elle assure le mieux la précision de la mesure par la réduction de l'importance de l'erreur, l'analyse de facettes, mise au point dans le modèle de généralisabilité, est mieux indiquée pour l'élaboration des épreuves à référence critérielle[4].

À propos de l'analyse de facettes, on peut néanmoins se poser quelques questions en vue de nouvelles recherches : « Du fait qu'elle prend en compte à la fois la variance d'interaction entre les faces D et I et la variance entre les niveaux de la face d'instrumentation, la procédure d'analyse d'items visant la minimisation de l'erreur absolue a-t-elle des effets moins prédictibles que ceux de la minimisation de l'erreur relative ? Quelle exploitation faut-il faire de l'analyse de facettes pour que chacune de ses perspectives (mesures relatives ou absolues) réalise mieux son objectif spécifique ? Par exemple, vaudrait-il mieux, pour la seconde, éliminer successivement des items pour leur contribution à la variance d'interaction entre faces D et I, d'une part, et pour leur contribution à la variance entre les niveaux des conditions d'observation, d'autre part, ce qui permettrait de tenir compte de l'importance relative de ces deux variances ? »

La démarche d'optimisation de la fidélité par la réduction de l'univers de généralisation préconisée dans cette étude ne va pas sans poser de problème. En effet, la sélection des conditions d'observation semble contredire le fondement même de la théorie de la généralisabilité qui suppose l'échantillonnage aléatoire des conditions d'observation. Ainsi, par exemple, si les questions sont sélectionnées soigneusement, elles ne peuvent pas logiquement être en même temps considérées comme choisies aléatoirement et donc comme sources de fluctuations d'échantillonnage.

Toutefois, il faut noter que l'optimisation de la mesure par la sélection des conditions d'observation peut être envisagée comme un moyen de mieux « cerner » l'univers de ces conditions, pour mieux en estimer les paramètres véritables, de la même façon que Tukey (1977) élimine de l'échantillon aléatoire observé les cas trop extrêmes qui perturbent l'estimation de la moyenne et de la variance, afin d'en obtenir des estimations plus « robustes ». À notre avis, les méthodes d'analyse d'items permettent d'accroître la *« robustesse »* de l'estimation des caractéristiques statistiques (moyennes des questions, des sujets, des phases, etc.) ainsi que celle des paramètres de généralisabilité, par l'élimination *après coup* des conditions d'observation atypiques ou trop extrêmes.

Mais de toute façon, les procédures d'analyse d'items peuvent conduire au moins à redéfinir l'univers des niveaux de la face d'instrumentation dans lequel on pourra *à l'avenir* échantillonner de façon assurée. La généralisabilité n'est-elle pas un rapport qui s'accroît si, d'une part, on élargit l'ensemble à différencier et si, d'autre part, on restreint l'ensemble sur lequel on veut généraliser?

NOTES

1. La variance totale d'un test est égale à la somme des termes de la matrice des variances-covariances entre les items tandis que la variance vraie est, quant à elle, fournie principalement par la somme des termes placés en dehors de la diagonale. Le rejet des items ayant une faible corrélation avec les autres, et donc avec le test total, accroît la valeur moyenne des termes hors diagonale et, par conséquent, la fidélité du test.
2. La même analyse s'effectue également sur la face de différenciation.
3. Voir à ce sujet Mokonzi (2001).
4. Les perspectives normative et critérielle de la mesure sont cependant applicables à n'importe quelle facette d'un plan d'observation.

RÉFÉRENCES

Bain, D. (à paraître dans ce numéro). Généralisabilité et séquences didactiques : illustration et défense d'un modèle à vocation édumétrique. *Mesure et évaluation en éducation.*

Bain, D. & Pini, G. (1996). *Pour évaluer vos évaluations. La généralisabilité : mode d'emploi.* Genève : Centre de recherches psychopédagogiques. Direction générale du cycle d'orientation.

Cardinet, J. (1986). *Évaluation scolaire et mesure*. Bruxelles : De Boeck.

Cardinet, J. & Tourneur, Y. (1985). *Assurer la mesure*. Berne : Peter Lang.

Mokonzi, G.B. (2001). *Exploitation de quelques méthodes de sélection des données pour l'optimisation des mesures en éducation : une étude comparative de la psychométrie classique et de la théorie de la généralisabilité*. Thèse de doctorat inédite, Université de Kisangani.

Tukey, J. (1977). *Exploratory data analysis.* Reading, Mass. : Addison-Wesley.