Mesure et évaluation en éducation

La détermination de l'unidimensionalité de l'ensemble des scores à un test

Jean-Guy Blais et Michel Laurier

Volume 20, numéro 1, 1997

URI : https://id.erudit.org/iderudit/1091387ar
DOI : https://doi.org/10.7202/1091387ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Blais, J.-G. & Laurier, M. (1997). La détermination de l'unidimensionalité de l'ensemble des scores à un test. *Mesure et évaluation en éducation*, *20*(1), 65–90. https://doi.org/10.7202/1091387ar

Résumé de l'article

Les objectifs de l'étude présentée dans ce document sont, d'une part, d'illustrer une démarche de détermination statistique de l'unidimensionalité d'un ensemble de scores à un test mettant l'accent sur la validité et, d'autre part, d'étudier l'efficacité de la proposition de Stout (1987 et 1990) en contrastant les résultats obtenus avec deux approches complémentaires : l'analyse de la structure des relations et l'analyse factorielle de Bock et al. (1985).

Mesure et évaluation en éducation
Vol. 20 n° 1, pages 65 à 90.

La détermination de l'unidimensionalité de l'ensemble des scores à un test

Jean-Guy Blais et Michel Laurier
Université de Montréal

Les objectifs de l'étude présentée dans ce document sont, d'une part, d'illustrer une démarche de détermination statistique de l'unidimensionalité d'un ensemble de scores à un test mettant l'accent sur la validité et, d'autre part, d'étudier l'efficacité de la proposition de Stout (1987 et 1990) en contrastant les résultats obtenus avec dceux approches complémentaires : l'analyse de la structure des relations et l'analyse factorielle de Bock et al. (1985).

(unidimensionalité, indépendance locale, validité)

The objectives of this study are twofold. First, to illustrate a way to technically determine tests scores unidimensionality with a focus on test validity and, secondly, to study Stout's proposal (1987 and 1990) in contrast with a structural equation solution and Bock's full information factor analysis *solution (1985).*

(unidimensionality, local independence, validity)

Introduction

Les efforts investis dans l'étude de l'unidimensionalité d'un test, d'un ensemble d'items ou d'un ensemble de scores se sont intensifiés depuis que les propositions de la théorie de la réponse à l'item (TRI) sont à l'avant-scène de la modélisation des scores à un test. En effet, étant donné que, parmi ces propositions, il s'en trouve une grande majorité dont la formulation mathématique explicite est unidimensionnelle, il est normal que l'on cherche à vérifier si cette caractéristique du modèle est réaliste dans les situations où on veut appliquer le modèle.

Les modèles de la théorie de la réponse à l'item les plus utilisés posent donc comme hypothèse de modélisation que la performance d'un candidat peut être décrite par une seule variable latente. Cette hypothèse n'est pas

nouvelle dans le domaine du développement des tests en éducation et en psychologie. Même si elle n'a pas toujours été formalisée comme dans la théorie de la réponse à l'item, l'unidimensionalité a toujours été une hypothèse fondamentale dans la théorie des tests où on affirme souvent que les scores possèdent une certaine pertinence dans la mesure où les items du test ne mesurent qu'un seul trait (Hattie et coll. 1996; McNemar, 1946, p. 298). Ainsi, si le test est composé d'items qui mesurent différents traits, on pense qu'il est alors difficile d'interpréter les scores, de les mettre en relation avec d'autres scores ou de mettre en relief les différences individuelles.

Cette hypothèse pourrait être passablement restrictive si l'on en exigeait la démonstration absolue car la probabilité que, dans une situation réelle de *testing*, il n'y ait vraiment qu'une seule et unique dimension semble plutôt faible (Goldstein, 1980; Humphreys, 1984). Dans une situation réelle, plusieurs éléments entrent en jeu qui influencent toujours, jusqu'à un certain point, la performance au test et donc les scores observés. Ces éléments sont associés autant aux items (individuellement et collectivement) qu'aux candidats et au contexte, et aux interactions entre chacun.

La théorie de la réponse à l'item fait partie d'une classe de modèles, au même titre que l'analyse des classes latentes et que l'analyse de la structure latente, qui tentent d'expliquer la covariation d'un nombre donné de variables latentes qui caractériseraient une population de candidats (McDonald, 1989). Les variables latentes du modèle représentent symboliquement différentes caractéristiques non observables des candidats, que l'on désigne comme des traits latents[1] (pour la définition de trait, voir Zuroff, 1986). Les scores au test représentent une manifestation de la mise en action de traits latents des candidats, dans des conditions données et en fonction de certaines tâches (les items). Même si, souvent, on attribue le qualificatif undimensionnel au test ou à l'ensemble d'items, c'est vraiment à l'ensemble des scores qu'il doit être attribué parce que ce sont ceux-ci qui constituent la trace de la rencontre entre les candidats et les items. De même, lors de la validation d'un test, l'accent est mis sur les scores (la mesure) en opposition au test (l'instrument) parce que les propriétés qui définissent une information adéquate sont des propriétés associées aux scores et non au test. La démonstration technique de l'undimensionalité d'un test repose donc, avant tout, sur l'étude de l'ensemble de scores qui est généré par la rencontre entre des tâches et des candidats, dans des conditions données.

Dans cette perspective, il semble qu'il soit beaucoup plus réaliste et pratique de considérer que l'hypothèse d'unidimensionalité est vérifiée lorsqu'on peut montrer qu'une dimension dominante explique ou est

responsable de la performance et des réponses des candidats (Humphreys, 1984). D'autre part, même si les scores peuvent permettre certaines vérifications techniques de l'hypothèse d'unidimensionalité, ils peuvent rester difficiles à interpréter conceptuellement si le trait visé par le test ne peut pas être défini clairement et sans ambiguïté, d'autant plus qu'il n'est pas observable directement. Ainsi, même si l'expression « dimension dominante » peut rendre la vie plus facile d'un point de vue technique, la définition du concept d'unidimensionalité, en fonction de variable latente unique, devrait être concrète et opérationnelle (Hambleton & Rovinelli, 1986).

C'est peut-être pour pallier ces difficultés d'interprétation que Reckase (1990) a rappelé qu'il existe une distinction entre le construit psychologique visé et les outils statistiques employés pour confirmer l'existence du construit. Ainsi, dans le cadre particulier de l'étude de l'unidimensionalité, il a proposé de faire une distinction entre l'unidimensionalité psychologique et l'unidimensionalité statistique. La caractéristique d'unidimensionalité psychologique du test fait référence à la définition conceptuelle ou théorique du trait mesuré par les items tandis que la caractéristique d'unidimensionalité statistique est en quelque sorte une proposition de définition opérationnelle de l'unidimensionalité psychologique. Cette distinction rapproche l'unidimensionalité psychologique du concept de validité de construit du test tel que favorisé par Messick (1989). Dans cette perspective, la présence de la dimension unique, non directement observable, ne peut être qu'inférée à partir d'une analyse fine du modèle conceptuel ou théorique qui contribue à la mise au point des items, d'une analyse du contenu des items, d'une analyse des liens que les scores au test entretiennent avec les scores à d'autres tests et d'une analyse des liens avec les objectifs présidant à la mise au point du test. Si le test est conçu pour mesurer une seule habileté, la démonstration de sa validité intègre donc la démonstration de son unidimensionalité psychologique.

Par ailleurs, il est possible que la méthode acceptée[2] de démonstration de la validité de construit d'un test qui vise à mesurer un seul trait, une seule habileté, suggère également de faire appel à des techniques d'analyse des données qui ont pour rôle de démontrer l'unidimensionalité statistique. Dans ce cas, la démonstration de l'unidimensionalité statistique sera concourante à la démonstration de l'unidimensionalité psychologique, mais elle ne pourra jamais lui être substituée. En effet, même si, en tant que proposition de définition opérationnelle, l'unidimensionalité statistique a souvent et traditionnellement fait office de substitut à l'unidimensionalité psychologique, la démonstration de l'unidimensionalité d'un test est une démarche

qui devrait se réaliser en ayant à l'esprit la priorité de la démonstration de la validité du test.

Dans ce qui suit, nous étudions l'efficacité de la proposition de Stout (1987, 1990) pour déterminer la dimensionalité des ensembles de scores à une version expérimentale d'un test de placement en français langue seconde; l'instrument contient trois sous-tests de 50 items avec quatre choix de réponse. L'intérêt de cette proposition est qu'elle ouvre une porte à la contribution de la réflexion du chercheur sur la validité de construit du test, permettant ainsi une confrontation entre la perspective statistiques et la perspective conceptuelle. De plus, la complémentarité de la proposition de Stout, de l'analyse de la structure des relations et de l'analyse factorielle de Bock et coll. (1985) sera également étudiée. La présentation des résultats sera précédée d'un tour d'horizon d'approches proposées pour « démontrer » l'unidimensionalité statistique des scores à un test et de l'introduction d'une proposition de définition formelle de l'unidimensionalité.

La détermination de l'unidimensionalité statistique

L'importance, dans le cadre de la théorie de la réponse à l'item, de pouvoir démontrer qu'une dimension dominante est responsable de la performance des candidats est accentuée par le fait qu'un certain nombre de simulations et d'études, avec des données réelles, ont démontré que les paramètres de différents modèles unidimensionnels sont mieux estimés lorsqu'il n'y a qu'une seule dimension présente dans les données (pour les simulations) ou dans la structure conceptuelle du test (pour les données réelles). On retrouve des problèmes d'estimation des paramètres associés à la dimension visée lorsque d'autres dimensions prennent plus d'importance (pour ces résultats, voir Blais, 1987; Doody-Bogan & Yen, 1983; Drasgow & Parsons, 1983; Greaud, 1988; Harrison, 1986; Kim & Stout, 1993; Reckase, 1979; Wang, 1988).

D'un autre point de vue, il existe également plusieurs propositions de modélisation multidimensionnelle qui pourraient être plus rentables pour représenter les données (la rentabilité d'une procédure étant évidemment liée à l'objectif poursuivi). Toutefois, l'utilité des modèles multidimensionnels reste encore à être illustrée à l'extérieur d'études où les données sont simulées, c'est-à-dire dans des situations où il y a une place importante qui doit être faite à l'interprétation. En effet, il reste difficile de commenter et d'apprécier le fonctionnement d'un item lorsqu'il y a, par exemple, trois

indices de difficulté et trois indices de discrimination, et que chaque candidat se voit attribuer un vecteur d'habileté de dimension trois. De plus, il existe plusieurs façons de définir et donc de simuler une structure multidimensionnelle. Est-ce que cela implique que l'on doive entrevoir la possibilité de l'existence de plusieurs types de multidimensionalité qui seraient déterminés par autant de procédures différentes? Pour ces différentes raisons, les applications multidimensionnelles se sont faites plutôt rares jusqu'à maintenant (mais, voir Luecht, 1996, et van der Linden, 1996) et la démonstration de l'unidimensionalité statistique des scores à un test demeure une préoccupation importante de la communauté des chercheurs en psychométrie.

Plusieurs suggestions ont été mises de l'avant pour élaborer une méthode statistique qui fournisse une définition opérationnelle efficiente de l'unidimensionalité statistique. L'approche, que l'on pourrait qualifier de « classique », fait appel aux procédures associées à l'analyse factorielle de Spearman (Zwick, 1987). Par exemple, McDonald (1981, pp. 14-15) conclut « qu'il est relativement raisonnable d'affirmer qu'un ensemble de n tests ou un ensemble de n items dichotomiques est unidimensionnel si et seulement si on peut lui ajuster un modèle factoriel non linéaire avec un facteur commun».

Hattie (1984, 1985) a produit une étude détaillée de certaines procédures statistiques ayant été suggérées pour déterminer si l'ensemble des scores à un test est unidimensionnel. Les différentes procédures recensées par Hattie peuvent être classées selon qu'elles utilisent les schémas de réponse, selon qu'elles sont issues de différentes théories des tests comme la théorie classique des tests (les indices de fidélité, de consistance interne et d'homogénéité) et la théorie de la réponse à l'item (les statistiques d'adéquation), ou selon leurs liens avec des techniques de réduction des données comme l'analyse factorielle (les valeurs propres, l'étude des résidus).

Ainsi, lorsque les items sont ordonnés selon leur degré de difficulté, les scores provenant d'un test unidimensionnel devraient permettre d'obtenir une « échelle » de Guttman (c'est-à-dire, une hiérarchie des items établie selon leur degré de difficulté et une hiérarchie des répondants formée d'après le nombre de bonnes réponses). Le degré d'adéquation entre les items et une échelle de Guttman peut être apprécié à l'aide de différents indices (Cliff, 1983), mais ceux-ci semblent plus utiles pour détecter des schémas de réponses anormaux que pour détecter la multidimensionalité (Hattie, 1984).

Les indices associés à la consistance interne d'un test, comme l'indice alpha de Cronbach ou encore les formules K-R 20, K-R 21 de Kuder et Richardson, prennent des valeurs moins élevées lorsqu'un test vise à mesurer des traits non corrélés. Par rapport aux items, des valeurs faibles des indices de discrimination correspondent habituellement soit à des items présentant des failles de conception ou à des items qui ne sont pas en lien avec le score total, c'est-à-dire des items qui potentiellement ne mesureraient pas le trait dominant.

L'analyse factorielle ou l'analyse en composante principale demandent d'abord d'appliquer ces procédures à la matrice des corrélations tétrachoriques. Ensuite, on peut observer la contribution de chacun des facteurs par le biais de la variance « expliquée » ou en examinant l'ordre de grandeur des premières valeurs propres. On peut également suivre les recommandations de McDonald (1981) et examiner les résidus résultant de l'ajustement d'un modèle non linéaire avec un facteur commun. Finalement, on peut aussi utiliser une procédure d'analyse factorielle basée sur la théorie de la réponse à l'item (*full-information factor analysis* de Bock et coll. 1985) qui fait appel à une version multidimensionnelle du modèle normal à trois paramètres (par exemple, Zwick, 1987). Cette procédure ne fait pas intervenir la matrice de corrélations et évite ainsi les problèmes associés à son utilisation avec des scores dichotomiques (Mislevy, 1986).

Dans le cadre de la théorie de la réponse à l'item, Bejar (1980) propose une procédure demandant de comparer les estimations des paramètres obtenues d'abord avec le test complet et ensuite avec des sous-ensembles d'items regroupés selon la pertinence du contenu. La procédure a été utilisée pour supporter l'hypothèse d'unidimensionalité dans des tests d'habileté langagière (Henning et coll., 1985), mais elle a donné de moins bons résultats lors de certaines simulations (Hambleton & Rovinelli, 1986).

De son côté, Hambleton (1989) s'inspire des résultats de l'étude de Hattie et recommande six procédures qu'il établit comme les plus prometteuses pour vérifier l'unidimensionalité. Quatre de ces procédures sont en lien avec l'esprit et la technique de l'analyse factorielle. Une autre procédure examine le lien entre l'unidimensionalité et l'hypothèse d'indépendance locale et la dernière procédure demande une double calibration des items à l'intérieur d'un modèle de la théorie de la réponse à l'item et la comparaison des valeurs des paramètres issues de chacune des calibrations.

Depuis les travaux de Hattie et les recommandations de Hambleton, d'autres procédures statistiques ont été proposées pour servir de définition

opérationnelle à l'unidimensionalité psychologique. Par exemple, on peut mentionner les travaux de Rozenbaum (1984) et Holland et Rozenbaum (1986), et plus récemment, la procédure développée par Stout (1987, 1990) qui a été améliorée par Nandakumar et Stout (1993) et la procédure de Chen et Thissen (1997). On peut également anticiper, même si on compte peu de travaux dans ce sens à l'heure actuelle, que des procédures impliquant la modélisation de la structure latente puissent servir à confirmer une structure donnée, qu'elle soit unidimensionnelle ou multidimensionnelle.

Le principe sous-jacent à la plupart des approches utilisées pour déterminer l'unidimensionalité statistique est celui de l'appréciation de la covariation des scores aux items du test. Mais, trop souvent, ce principe souffre de l'absence d'une définition formelle de ce qu'est l'unidimensionalité statistique. Lorsqu'on replace le concept de dimensionalité dans le contexte des théories du trait latent, il existe une définition formelle de ce qu'est la dimensionalité. Cette voie a été empruntée par Stout (1987; 1990), par Holland et Rozenbaum (1986) et par Chen et Thissen (1997). Elle demande de faire la jonction avec le concept d'indépendance locale, tel que l'a suggéré McDonald (1981).

Le nombre de dimensions et l'indépendance locale

Pour Lord et Novick (1968, pp. 531-541) et McDonald (1981), la notion de dimensionalité est régie par le principe d'indépendance locale. Le nombre de dimensions k d'un ensemble de n mesures est le nombre minimal de traits latents produisant des réponses indépendantes pour ces n mesures. Il y a indépendance locale si, étant donné un ensemble de traits latents, n mesures sont indépendantes en probabilité dans une sous-population de candidats se situant au même endroit sur le continuum des valeurs prises par chaque trait latent.

Soit $\mathbf{U}_n = (U_1, U_2, \ldots, U_n)$, le vecteur des variables identifiant les scores aux items (par exemple, $U_i = 0$ ou 1) et Θ le vecteur des traits latents $(\theta_1, \theta_2, \ldots, \theta_k)$. La théorie de la réponse à l'item introduit une fonction

$$P_i(\theta) = [U_i = u_i \mid \Theta = \theta]$$

décrivant pour chaque item la probabilité qu'un candidat choisi aléatoirement dans un groupe de candidats d'habileté $\Theta = \theta$ réussisse l'item i et se voit attribuer un score $U_i = u_i$ (Stout, 1990).

La condition d'indépendance locale exige que pour chaque schéma de réponses (u_1, u_2,..., u_n) et pour chacune des valeurs θ de Θ :

$$P_i\left[U_1 = u_1, U_2 = u_2,\ldots,U_n = u_n \mid \Theta = \theta\right] = \prod_{i=1}^{n} P_i\left[U_i = u_i \mid \Theta = \theta\right]$$

Le nombre de dimensions d de l'ensemble des scores sera la dimensionalité minimale requise du vecteur Θ pour produire l'indépendance des fonctions $P_i(\theta)$.

Cette condition à l'indépendance locale est très stricte parce que non seulement elle exige que les covariances entre les scores soient nulles, mais également que tous les moments supérieurs soient des produits des moments univariés (Hattie et coll, 1996; McDonald, 1981).

Une définition moins stricte demanderait de vérifier uniquement si les covariances entre les scores sont nulles (McDonald, 1981). On exercerait ainsi une distinction entre une condition stricte d'indépendance locale, la condition forte, et une condition moins stricte, la condition faible.

La condition faible d'indépendance locale pourrait s'exprimer, comme le propose Stout (1990), en définissant une condition d'« indépendance essentielle » qui tiendrait si :

$$\lim_{n \to \infty} \frac{\sum_{1 \le i < j \le n} \left|\mathrm{Cov}(U_i, U_j \mid \Theta = \theta)\right|}{\binom{n}{2}} \to 0$$

La condition tient donc si la moyenne des covariances entre toutes les paires d'items tend vers zéro lorsque le nombre d'items tend vers l'infini. Le nombre de dimensions essentielles d_E (que Stout appelle l'unidimensionalité essentielle) de l'ensemble des scores au test serait donc le nombre minimal de trait θ nécessaire à la réalisation de cette expression du principe faible d'indépendance locale.

On remarque un certain nombre de choses dans ce rapprochement entre la dimensionalité et l'indépendance locale, d'une part, et la proposition de Stout, d'autre part. D'abord, la définition de l'indépendance essentielle est élaborée en fonction d'un vecteur Θ de traits latents. On rejoint ainsi l'idée de la présence d'une dimension dominante puisque même si plusieurs dimensions contribuent à la production des réponses observées, cela

n'empêche nullement la réalisation de la condition d'indépendance essentielle. Dans la proposition de Stout, on ne démontre pas qu'il n'y a qu'une seule dimension, mais plutôt qu'il existe une représentation des scores par un modèle unidimensionnel monotone pour lequel la condition faible d'indépendance locale tient (Stout, 1987, 1990). L'approche de Stout est élaborée dans le contexte d'un nombre d'items infini où les propriétés des estimateurs statistiques sont asymptotiques. Dans des situations où le nombre d'items est réduit, il n'est pas sûr que les propriétés tiennent (dans cette direction, voir de Champlain et Gessaroli, 1991). Finalement, si l'ensemble des scores est unidimensionnel, la représentation de θ, l'échelle de l'habileté, est arbitraire parce que toute transformation strictement monotone de θ produit le même modèle unidimensionnel. Le score vrai (le nombre espéré de bonnes réponses) constituant une transformation monotone de θ (Lord, 1980, p. 46), un estimateur consistant du score vrai peut être utilisé pour démontrer la condition d'indépendance locale faible. Comme le nombre de bonnes réponses constitue un tel type d'estimateur, c'est cette quantité qui est substituée pour calculer les estimations des covariances entre les paires d'items dans l'opérationnalisation que propose Stout.

L'opérationnalisation de Stout (1987 et 1990)

Stout (1987) a proposé une définition empirique de l'unidimensionalité essentielle sur laquelle repose un test statistique de l'hypothèse d'unidimensionalité : $H_0 : d_E = 1$; $H_1 : d_E > 1$, où d_E est l'unidimensionalité essentielle de l'ensemble des scores au test. La proposition a par la suite été améliorée par Nandakumar et Stout (1993). Cette proposition vise exclusivement les ensembles de scores dichotomiques constitués de 0 et de 1.

La procédure menant au test de l'hypothèse ci-dessus a été décrite, notamment, par Nandakumar et Stout (1993) et elle se déroule selon les quatre étapes suivantes :

1. On sélectionne les *M* items qui feront partie du premier sous-test de vérification, STV1. Pour des considérations de robustesse de l'estimation, le nombre d'items composant STV1 ne devrait pas dépasser le quart du nombre total d'items. Deux stratégies sont suggérées pour constituer STV1 : a) une analyse conceptuelle de l'ensemble des items par un ou des experts pour produire une sélection d'un sous-ensemble d'items le plus unidimensionnel possible; b) une analyse en composantes principales de la matrice des corrélations tétrachoriques, où ce sont les *M* items ayant les saturations les plus élevées sur le deuxième facteur (avant rotation) qui sont sélectionnés pour faire partie de STV1 (par

ailleurs, Roussos et coll., 1993, cités dans Hattie et coll., 1996, ont également proposé d'utiliser l'analyse en grappe pour créer STV1).

2. Un second ensemble de M items est sélectionné à partir des items restants, de façon que la difficulté et la dimensionalité de l'ensemble d'items ressemblent à ce qu'on retrouve pour STV1. Cet ensemble constitue le deuxième sous-test de vérification, STV2. Ce sous-test sera utilisé pour apporter une correction à la statistique issue de STV1.

3. Les items non utilisés pour STV1 et STV2, $N - 2M$ items, forment le sous-test de répartition, STR. Les scores au sous-test de répartition servent à regrouper les candidats selon le résultat obtenu. Ainsi, en excluant les candidats qui n'ont que des bonnes ou des mauvaises réponses, le sous-test de répartition permet de former au plus $N-2M-2 = R$ regroupements. Pour conserver les propriétés asymptotiques de la statistique, il est suggéré de former des regroupements d'au moins vingt candidats.

4. On estime la variance des scores pour chacun des sous-tests de vérification, $\hat{\sigma}_r^2$, et la variance unidimensionnelle, $\hat{\sigma}_{U,r}^2$, pour chacun des r regroupements de candidats produits par le sous-test de répartition[3]. On calcule une statistique T_i pour chaque sous-test de vérification où = $S_r^2 = Var(\hat{\sigma}_r^2 - \hat{\sigma}_{U,r}^2)$.

$$T_i = \frac{1}{R^{1/2}} \sum_{r=1}^{R} \left[\frac{\hat{\sigma}_r^2 - \hat{\sigma}_{U,r}^2}{S_r} \right]$$

Finalement, on calcule la statistique $T = \frac{T_1 - T_2}{\sqrt{2}}$ et on vérifie l'hypothèse

$d_E = 1$ en profitant du fait que la distribution de T est asymptotiquement normale avec une moyenne 0 et une variance 1 (Stout, 1987).

Essentiellement, la procédure de Stout vérifie le degré de proximité entre un modèle essentiellement unidimensionnel et le modèle qui a généré les scores observés. La statistique T_1 est une information sur le degré de multidimensionalité que l'on retrouve localement pour le regroupement r. Elle est sensible à la multidimensionalité et au biais de l'estimation. La statistique T_2 est calculée à partir d'un ensemble d'items, STV2, que l'on considère équivalent à l'ensemble STV1, et elle est utilisée pour corriger le biais d'estimation de la statistique T_1 (Hattie et coll., 1996, p. 3).

Une étude de la dimensionalité des ensembles de scores provenant d'un test de placement en français langue seconde

La procédure de Stout est disponible avec un programme informatique, *DIMTEST* (Stout et coll., 1991) fonctionnant sous le système *DOS* de la compagnie *Microsoft*. En utilisant ce programme, nous avons mis en parallèle la procédure de Stout et deux autres approches d'appréciation de l'unidimensionalité statistique. Ces approches et les outils informatiques qui les supportent sont : a) la modélisation de la structure des relations entre les scores en utilisant le logiciel *LISREL* (Jöreskog & Sörbom, 1983); b) l'analyse factorielle de Bock et coll. (1985) avec le logiciel *TESTFACT* (Wilson et coll. 1991).

Dans chacune de ces trois approches, il y a un certain espace pour intégrer des données provenant du jugement d'experts concernant la structure conceptuelle de l'ensemble d'items. Ainsi, en contrastant les résultats des différentes analyses et en mettant en perspective les connaissances des chercheurs, on pourra apprécier plus globalement l'efficacité de la proposition de Stout.

La procédure de Stout laisse la liberté au chercheur de préciser lui-même quels sont les items qui composent les sous-tests de vérification et qui doivent former des ensembles unidimensionnels. De plus, cette procédure est la seule qui propose un test statistique pour éprouver l'unidimensionalité d'un ensemble de scores. Les procédures de modélisation de la structure des relations exigent également une intervention du chercheur. Celui-ci doit d'abord proposer une structure de relations et vérifier ensuite avec les scores si cette structure se confirme. L'analyse factorielle de Bock va dans le même sens puisqu'elle demande une interprétation conceptuelle des regroupements d'items en facteurs. Ainsi donc, dans ces trois propositions, le chercheur va au-delà de la simple utilisation d'un indicateur statistique et met à contribution sa réflexion sur la validité de construit et sur l'unidimensionalité psychologique du test.

Les ensembles de scores avec lesquels l'étude a été menée proviennent des réponses à une version expérimentale d'un test de placement en français langue seconde. Le test a été administré à des étudiants canadiens-anglais de différents collèges et universités, inscrits à des cours d'été de français dans le cadre d'un programme de bourse pour l'apprentissage d'une des langues officielles du Canada. Le test est divisé en trois sous-tests de 50 items chacun.

Pour le premier sous-test, les étudiants doivent lire un court paragraphe d'environ trente mots et répondre à des questions de compréhension à réponse choisie. Le deuxième sous-test demande aux étudiants de choisir, dans un ensemble de quatre expressions grammaticales, celle qui est la plus appropriée dans une situation de la vie courante (par exemple, comment féliciter une amie pour ses succès). Le troisième sous-test consiste en des phrases lacunaires où les réponses recherchées sollicitent des compétences en grammaire et en vocabulaire.

Étant donné les exigences du programme, le groupe des participants à ces cours était relativement homogène sur le plan de l'âge, de la scolarité, du statut socio-économique et des antécédents linguistiques[4]. Les ensembles de scores sont constitués d'un noyau de réponses de 348 étudiants qui ont répondu à l'ensemble des questions du test. Pour les sous-tests, ce nombre initial a pu être augmenté à 694 étudiants pour le premier sous-test, à 681 étudiants pour le deuxième sous-test et à 661 étudiants pour le troisième sous-test. Les trois procédures ont été appliquées soit à l'ensemble des scores du test complet (N = 348), soit aux ensembles provenant des sous-tests. Ainsi, on visait à déterminer si le test dans son ensemble est unidimensionnel et si chaque sous-test est unidimensionnel. La décision prise pourrait avoir un impact non négligeable sur l'utilisation de ce test parce qu'un utilisateur pourrait se voir contraint de travailler avec un, deux ou trois scores différents et peut-être même avec plus de trois scores si on en arrivait à la conclusion que certains sous-tests sont multidimensionnels.

La modélisation de la structure des relations

L'objectif de la modélisation de la structure des relations était de confirmer la pertinence de la division en trois sous-tests. Pour ce faire, chaque sous-test a d'abord été divisé en deux parties égales en attribuant aléatoirement les items pairs et impairs à l'une ou l'autre des deux parties. Une solution avec un facteur unique et une solution avec trois facteurs ont ensuite été mises à l'épreuve.

Le tableau 1 montre que les corrélations entre les différentes parties des sous-tests sont relativement élevées. Ces résultats pourraient nous amener à penser qu'un modèle avec un seul facteur serait acceptable si ce n'était du fait que l'adéquation de ce modèle aux données est plutôt faible. Le modèle avec trois facteurs corrélés a donc été ajusté aux données. La solution est apparue satisfaisante et la représentation structurelle de ce modèle est présentée à la figure 1.

Tableau 1

Corrélations entre les parties des sous-tests

	Sous-test 1		Sous-test 2		Sous-test 3	
	Pairs	Impairs	Pairs	Impairs	Pairs	Impairs
Sous-test 1 - Pairs	1,00	0,99	0,86	0,86	0,87	0,87
Sous-test 1 - Imp.	0,99	1,00	0,85	0,85	0,87	0,87
Sous-test 2 - Pairs	0,86	0,85	1,00	0,99	0,81	0,81
Sous-test 2 - Imp.	0,86	0,85	0,99	1,00	0,81	0,81
Sous-test 3 - Pairs	0,87	0,87	0,81	0,81	1,00	0,99
Sous-test 3 - Imp.	0,87	0,87	0,81	0,81	0,99	1,00

Les variables x^i définissent les réponses aux items et les scores correspondants et elles comportent une partie d'erreur de mesure δ^i. Les variables latentes ξ^i sont en relation λ^{ij} avec les scores et les variables latentes sont en corrélation ϕ^{ij}. Dans ce modèle, chaque variable latente est en relation avec un seul sous-test. Comme l'adéquation statistique du modèle aux données est satisfaisante, cette modélisation supporterait en quelque sorte la décision de séparer le test de 150 items en trois sous-test de 50 items.

Cette analyse ne nous donne aucune information sur la dimensionalité des sous-tests, mais il est clair que si ceux-ci mesurent plus d'une habileté, l'analyse de l'adéquation du modèle ajusté nous informe que la structure interne du test n'en est que peu affectée. On pourrait pousser plus à fond l'étude en subdivisant les sous-tests selon des construits appropriés plutôt qu'en utilisant une division arbitraire pairs/impairs. Il n'en reste pas moins qu'une modélisation simple semble adéquate et qu'elle permet une utilisation satisfaisante des scores pour l'objectif poursuivi.

L'analyse factorielle de Bock et coll. (1985)

La procédure d'analyse factorielle de Bock et coll. (1985) met à profit la stratégie d'estimation du maximum de vraisemblance marginale et évite ainsi les problèmes associés à l'utilisation de la matrice des corrélations entre des scores dichotomiques (Mislevy, 1986). La procédure a été appliquée à l'ensemble de scores provenant du test complet et à chacun des ensembles provenant des sous-tests.

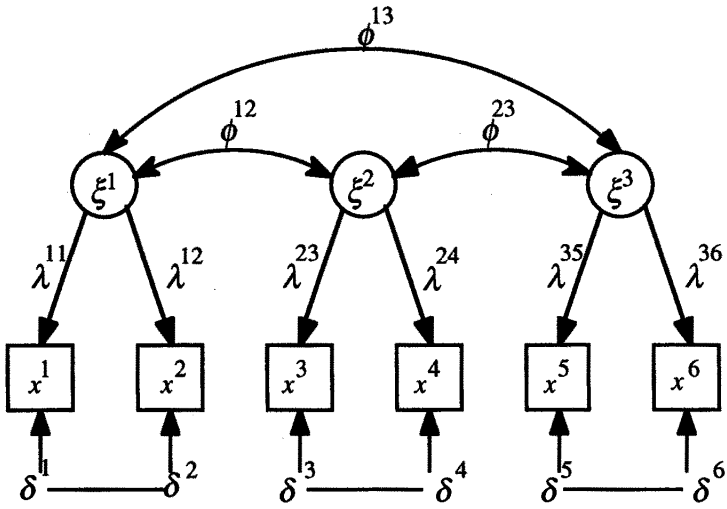

Figure 1 - Représentation du modèle avec trois facteurs corrélés

Pour le test complet, une solution de *TESTFACT* à trois facteurs indique que le premier facteur compte pour 25 % de la variance observée et que les deuxième et troisième facteurs comptent respectivement pour 2,4 % et 1,4 % de la variance observée. La corrélation entre les facteurs est relativement élevée : 0,60 entre le premier et le deuxième facteur; 0,68 entre le premier et le troisième facteur; 0,49 entre le deuxième et le troisième facteur. Un examen des saturations amène à constater que les 50 premiers items (le sous-test 1) sont surtout associés au premier facteur. L'habileté en lecture pourrait donc être posée comme une facette importante de l'habileté générale. Ce premier facteur retient aussi des items du troisième sous-test qui sont aussi clairement associés au troisième facteur. Ce troisième sous-test intégrerait une exigence d'habileté en lecture et une exigence spécifique d'habileté en grammaire. La situation quant à la nature du deuxième facteur est moins claire. Il n'y a que quelques items avec des saturations élevées et ces items semblent être associés à une habileté de raisonnement plutôt qu'à une habileté langagière. Pour les items du deuxième sous-test, les saturations sont généralement peu élevées, ce qui pourrait indiquer que ce sous-test est moins cohérent ou mesure plusieurs habiletés.

Pour l'analyse par sous-test, une solution à trois facteurs de *TESTFACT* pour le premier sous-test indique la présence d'un premier facteur dominant comptant pour 33 % de la variance observée. Avec une solution à deux facteurs, ce pourcentage augmente très légèrement pour atteindre 34 % avec un deuxième facteur à 2,3 %. Après une rotation oblique, la corrélation entre les deux facteurs est de 0,70. En examinant les saturations pour ces deux facteurs, on peut conclure que ces deux facteurs font ressortir deux habiletés cognitives différentes mais fortement corrélées, soit une habileté (dominante) à reformuler une information dans une deuxième langue et une habileté (secondaire) à faire des inférences à partir d'une information donnée. Pour le deuxième sous-test, le pourcentage de variance expliquée par le premier facteur d'une solution à trois facteurs est de 24 %, avec seulement six items dont les saturations sur le premier facteur sont élevées. De plus, on retrouve onze items dont les saturations sont faibles simultanément pour les trois facteurs. Pour le troisième sous-test, une solution à deux facteurs ne confirme pas la présence d'une distinction entre les compétences en grammaire et les compétences en vocabulaire. De surcroît, une solution à trois facteurs augmente de 23 % à 28 % le pourcentage de la variance expliquée que l'on peut associer au premier facteur.

On retrouverait donc après ces analyses sur les ensembles de scores provenant du test complet et des trois sous-tests : a) un premier sous-test où il y a un premier facteur nettement dominant et où les deux premiers facteurs sont fortement corrélés; b) un deuxième sous-test qui serait multidimensionnel, mais sans que nous ne puissions préciser tout à fait pourquoi; c) un troisième sous-test qui serait multidimensionnel et où la distinction que l'on établit entre compétences en grammaire et compétences en vocabulaire trouve sa justification dans l'analyse factorielle des scores issus du test complet mais pas dans l'analyse des scores au troisième sous-test.

L'approche non paramétrique de Stout

L'approche non paramétrique de Stout consiste en la production d'une statistique *T* dont la distribution asymptotique est une loi de probabilité normale de moyenne zéro et de variance un. Pour produire cette statistique, l'utilisateur du programme *DIMTEST* doit d'abord sélectionner un premier sous-test de vérification, STV1, contenant environ le quart des items. Comme nous l'avons déjà mentionné, cette sélection peut se faire de deux manières : à la suite des résultats d'une analyse factorielle ou à la suite des résultats d'une analyse conceptuelle de la part du chercheur. Ces deux façons de faire ont été mises à contribution pour analyser la dimensionalité de chacun des trois sous-tests.

Le tableau 2 présente les résultats obtenus pour chaque sous-test en utilisant la procédure d'analyse factorielle et la procédure de regroupements des items par un expert. Les items de chaque sous-test ont pu être regroupés en deux domaines A et B [5].

- Sous-test 1 : Domaine A : Habileté de reformulation (18 items)
  Domaine B : Habileté à réaliser une inférence (28 items)

- Sous-test 2 : Domaine A : Compétence lexicale (29 items)
  Domaine B : Compétence sociolinguistique (21 items)

- Sous-test 3 : Domaine A : Connaissance du vocabulaire (22 items)
  Domaine B : Connaissance des règles grammaticales (28 items)

Pour chaque domaine d'un sous-test, une sélection de 12 items (environ le quart des 50 items) a été effectuée pour satisfaire aux recommandations de Nandakumar et Stout (1993). Il faut souligner que le processus de sélection des items pour la constitution de STV1 peut poser des difficultés lorsqu'il faut choisir des items, pour former cet ensemble, dans un bassin qui dépasse largement le nombre suggéré (soit le quart des items). Certaines stratégies de sélection peuvent être plus intéressantes que d'autres. Ce problème sera traité plus à fond dans la partie présentant la discussion des résultats.

Comme on peut le constater à la première ligne du tableau 2, la procédure d'analyse factorielle construit des sous-tests de vérification pour lesquels T n'est jamais statistiquement significative. Ainsi, cette procédure nous amènerait à conclure que les trois sous-tests sont unidimensionnels. Dans le même tableau, aux lignes deux et trois, on constate que le regroupement des items en domaines A et B produit une statistique T qui n'est pas statistiquement significative pour les deux regroupements du premier sous-test, mais qui est statistiquement significative pour les deux regroupements des deuxième et troisième sous-tests. Les deux procédures de constitution des sous-tests de vérification ne mènent donc pas à des constatations tout à fait convergentes quant à l'unidimensionalité des ensembles des scores issus des sous-tests.

Discussion

Les résultats des analyses effectuées sont résumés dans le tableau 3 et nous amènent d'abord à constater qu'il y a certaines convergences entre les résultats, mais également que des approches utilisées isolément pourraient mener à différentes décisions concernant l'unidimensionalité d'un ensemble de scores.

Tableau 2

Les valeurs de la statistique *T* obtenues avec *DIMTEST*

	Sous-test 1	Sous-test 2	Sous-test 3
Analyse fact.	-1,50	0,28	-0,79
Domaine A	0,78	1,87*	1,71*
Domaine B	0,59	1,64*	2,76*

* *T* statistiquement significative avec $\alpha = 0,05$.

Ainsi, après l'analyse de la structure des relations avec le logiciel *LISREL*, on pourrait décider de scinder le score total en trois sous-scores, suivant alors la structure élaborée *a priori* par le concepteur du test. Cette décision s'appuierait sur le fait que la modélisation en trois sous-tests s'ajuste suffisamment bien aux données pour les objectifs poursuivis et qu'une modélisation plus complexe n'ajouterait pas substantiellement à la capacité du test à classer les étudiants dans les catégories appropriées.

Ensuite, l'analyse factorielle de Bock avec le logiciel *TESTFACT* sur le test complet pourrait nous amener à conclure qu'il existe deux dimensions plus importantes. Le premier sous-test constituerait la dimension dominante et le troisième sous-test une dimension secondaire où on retrouverait deux composantes. Mais l'existence d'une dimension associée directement aux scores du deuxième sous-test ne pourrait pas être confirmée.

Pour ce qui est de l'analyse avec les scores de chacun des sous-tests, on pourrait conclure à partir des résultats de *TESTFACT* que le premier ensemble de scores est unidimensionnel et que les deux autres sont multidimensionnels. Cependant, du point de vue conceptuel et étant donné les items qui possèdent les saturations les plus élevées, il serait difficile de déterminer la nature de

la multidimensionalité que l'on retrouve dans ces deux sous-tests. On pourrait également conclure, à partir des résultats d'une analyse avec *DIMTEST* qui intégrerait le jugement d'un expert, que le premier ensemble de scores est unidimensionnel et que les deuxième et troisième sont multidimensionnels. On peut ajouter que les deux façons de constituer les sous-tests de vérification avec *DIMTEST* ne produisent pas les mêmes résultats.

Tableau 3

Comparaison des décisions quant à la dimensionalité des ensembles de scores

	Sous-test 1	Sous-test 2	Sous-test 3	Test complet
Structure des relations (LISREL)	?	?	?	$d = 3$
Anal. factorielle (TESTFAC)	$d = 1$	$d > 1$	$d > 1$	$d = 2, 3$
DIMTEST (Anal. factorielle)	$d = 1$	$d = 1$	$d = 1$	?
DIMTEST (Expert)	$d = 1$	$d > 1$	$d > 1$	?

Quelles sont donc les conclusions finales à tirer de cette étude de l'unidimensionalité d'un test de placement en français langue seconde?

D'abord, il y a une certaine convergence des résultats amenant à conclure à la présence d'une dimension dominante pour l'ensemble de scores du premier sous-test. Ensuite, les ensembles de scores associés aux deuxième et troisième sous-tests apparaissent multidimensionnels, mais cette multidimensionalité est difficile à interpréter avec le cadre conceptuel qui a présidé à la mise au point du test et des sous-tests. Ainsi, même si les différentes analyses suggèrent la présence de plus d'une dimension pour ces deux sous-tests, il serait prématuré de les scinder en parties différentes à partir d'une règle uniquement technique dont l'application ne mène pas à des résultats faisant écho à la validité de construit des tests. Des études supplémentaires sont à prévoir dans cette direction.

D'autre part, la stratégie suivie dans cette étude est particulière au contexte et les outils techniques mis à contribution ne sont pas exempts de problèmes spécifiques. Dans n'importe quelle étude, la stratégie suivie est tributaire des conditions pratiques et de la profondeur de l'élaboration conceptuelle qui a présidé à la mise au point des tests. Dans le même sens, les outils privilégiés dans les études de ce type sont d'abord ceux qui sont disponibles à une époque donnée et, très souvent, pour lesquels on retrouve une procédure d'utilisation suffisamment conviviale[6]. Ces outils demeurent des propositions dont les propriétés et les comportements dans des situations variées doivent faire l'objet d'études élaborées jusqu'à ce qu'un consensus plus ou moins grand s'établisse quant aux conditions prévalant à une utilisation pertinente. Sinon, elles restent des propositions marginales survivant surtout dans les revues spécialisées, mais peu dans la pratique.

La stratégie que nous avons suivi pour déterminer l'unidimensionalité des ensembles de scores est ainsi fortement teintée du désir de valider des structures déjà présentes et de ne pas se fier uniquement à des indicateurs statistiques de l'unidimensionalité. Qui plus est, lorsque de l'analyse statistique émergent des résultats non interprétables à partir du cadre conceptuel, c'est ce dernier que nous décidons de privilégier plutôt que les indicateurs statistiques. Ainsi, même si deux des trois ensembles de scores peuvent être techniquement déclarés multidimensionnels, nous croyons que les scores des deux sous-tests visés ne devraient pas être scindés parce que l'interprétation de ces nouveaux scores n'est pas ancrée dans le cadre conceptuel.

De plus, les techniques mises en œuvre pour déterminer la dimensionalité statistique ne sont pas non plus à l'abri de problèmes de conception ou d'interprétation. Par exemple, une règle non écrite recommande qu'il est plus prudent, pour des raisons de robustesse de l'estimation des cœfficients de corrélation, d'utiliser une procédure d'analyse factorielle uniquement si le nombre de sujets est dix fois plus élevé que le nombre d'items. Même si la procédure d'analyse factorielle de Bock n'utilise pas de matrice de corrélations, on peut se demander si elle est aussi efficace peu importe le ratio nombre de sujets/nombre d'items.

D'autre part, on le soulignait plus haut, la constitution des sous-tests de vérification, dans l'approche de Stout à partir du jugement d'un expert, ne se fait pas aussi automatiquement que la procédure semble le suggérer. Il y a des décisions à prendre qui peuvent influencer la conclusion quant au statut de la dimensionalité de l'ensemble des scores. Par exemple, comme les sous-tests contiennent chacun 50 items, les sous-tests de vérification devraient intégrer

environ 12 items. Cette condition d'utilisation rend le test statistique plus robuste, mais la plupart du temps, elle impose au chercheur de faire le choix de ces items parmi un ensemble plus grand que douze. En effet, dans une situation réelle, l'expert réussira probablement à diviser l'ensemble des items du test en deux ou trois ensembles relativement homogènes. Il pourrait ainsi diviser un test de 50 items en trois sous-tests de 25, 15 et 10 items respectivement. Lequel de ces sous-tests devrait être mis à contribution et lesquels douze items devrait-on choisir parmi les ensembles où on en retrouve plus de douze? Est-ce qu'un choix aléatoire conviendrait et est-ce que tous les choix aléatoires donneraient des réponses indiquant des tendances similaires?

Des résultats de recherche préliminaires (Blais & Laurier, 1995, 1997) montrent que les réponses à ces questions stimulent une étude plus poussée des comportements de la statistique *T* proposée par Stout. Par exemple, on peut, dans un premier temps, se demander combien de fois la statistique *T* serait statistiquement significative si l'on créait 500 échantillons aléatoires de 12 items parmi les 28 items sélectionnés par l'expert pour faire partie du domaine B (connaissance des règles grammaticales) du sous-test trois. On pourrait ainsi estimer le degré de variabilité potentielle de la décision à prendre. Nous avons trouvé que la statistique *T* était statistiquement significative dans 18 % des cas, lorsque c'était le regroupement B qui faisait l'objet de l'échantillonnage des douze items, et dans 8 % des cas si c'était le regroupement A. Ainsi donc, près d'une fois sur cinq on déclarerait l'ensemble de scores multidimensionnel lorsque le choix des douze items était fait à partir du regroupement B et une fois sur dix si ce choix était fait à partir du regroupement A. L'étude plus poussée des propriétés de la statistique *T* étant donné les caractéristiques des items et des scores seraient donc une avenue de recherche à envisager qui rendrait grandement service à tous les futurs utilisateurs du logiciel *DIMTEST*.

Les résultats de Hattie et coll. (1996) nous éclairent également sur d'autres caractéristiques de la procédure proposée par Stout. D'après les résultats d'une simulation, la procédure d'analyse factorielle qui laisse *DIMTEST* choisir les items doit être employée uniquement si les matrices de corrélations tétrachoriques sont définies positives. De plus, lorsque le modèle utilisé pour simuler des ensembles de scores multidimensionnels est un modèle non compensatoire, les performances de la statistique *T* ne sont pas stables et le taux de détection de la multidimensionalité laisse à désirer. Par ailleurs, Nandakumar et Stout (1993) ont observé, également dans une simulation avec un modèle de la théorie de la réponse à l'item, que la

statistique *T* n'est pas très fiable lorsque les valeurs attribuées aux paramètres de discrimination et de « pseudo-chance » sont élevées. Finalement, une autre simulation par de Champlain et Gessaroli (1991) a illustré que la précision de la statistique *T* est influencée par la longueur du test et par la taille du groupe de répondants. Leurs résultats les amènent à recommander l'utilisation de *DIMTEST* avec des groupes de plus de 500 répondants et des tests d'au moins 25 items.

Conclusion

Le travail à la base de cet article avait comme objectif, d'une part, d'illustrer une démarche de détermination de l'unidimensionalité d'un ensemble de scores à un test et, d'autre part, d'y déterminer l'efficacité de la proposition de Stout (1987, 1990) en contrastant les résultats avec deux approches complémentaires : l'analyse de la structure des relations et l'analyse factorielle de Bock et coll. (1985). L'intérêt de la proposition de Stout réside d'abord dans le fait qu'elle amarre le concept de multidimensionalité au concept d'indépendance locale et, ensuite, que son opérationnalisation ouvre la porte à une contribution de la réflexion du chercheur sur la validité de construit du test, permettant ainsi une confrontation entre les perspectives statistiques et conceptuelles.

Les résultats nous amènent à réitérer le fait que la détermination de l'unidimensionalité de l'ensemble des scores à un test ne devrait pas être une décision dichotomique « oui/non » prise à la suite de l'obtention d'un résultat statistiquement significatif. Des techniques différentes pourraient produire des résultats différents et quelquefois contradictoires. Est-ce que cela signifie que le concept est défini selon la technique mise en œuvre pour réaliser l'analyse des données et qu'ainsi, il ne s'agit plus simplement de choisir la meilleure technique mais également de déterminer quel type d'unidimensionalité nous intéresse? L'importance de l'unidimensionalité est plutôt une question de degré. On pourrait tolérer des écarts plus ou moins importants selon le type de décision à prendre. La recherche de l'unidimensionalité ne devrait surtout pas restreindre le concepteur d'un test quant à la nature et à la diversité des tâches élaborées. L'étude de la dimensionalité est une partie de la démarche de la validation du construit et, donc, est un processus à long terme dans la recherche d'une proposition adéquate mais perfectible.

En effet, il faut distinguer le construit visé par le test, le trait mis en œuvre par les candidats pour réagir aux stimuli que sont les items du test, les réponses fournies aux stimuli et les scores que l'on attribue à ces réponses. Le construit provient d'une conceptualisation de la part de celui qui élabore l'instrument. Le concepteur veut mesurer quelque chose qui est soit une construction de l'imagination, non observable directement, soit un phénomène directement observable lorsque, par exemple, il prend la forme d'un comportement précis. La définition du construit à mesurer reste, pour l'instant, toujours reliée autant au savoir et à la méthodologie du moment qu'à la culture et à aux valeurs environnantes. En ce sens, la démonstration de la validité du construit à mesurer est toujours provisoire, c'est un processus continu en constant renouvellement selon les contributions des différentes influences théoriques et techniques du moment.

Ainsi, les scores, à partir desquels se feront toutes les analyses visant à démontrer la validité et, le cas écheant, l'unidimensionalité, sont dépendants de la mise en action d'un processus complexe où le ou les concepteurs, les tâches, les candidats, les réponses, le ou les juges et les contextes contribuent chacun à leur manière à influencer la production des scores observés et, par le fait même, à influencer les décisions à prendre. Il semble donc impératif, lorsqu'on étudie l'unidimensionalité d'un test, d'une part de subordonner l'approche statistique à l'approche conceptuelle et, d'autre part, de multiplier les façons de regarder statistiquement l'unidimensionalité. En l'absence d'un consensus quant à la définition technique de l'unidimensionalité, une « triangulation » statistique peut ainsi permettre de saisir des facettes différentes du test, qui sont confrontées au cadre conceptuel et qui ouvrent la porte à des prises de décision plus nuancées parce que basées sur une plus grande variété de données.

NOTES

1. Dans le vocabulaire du domaine cognitif, on utilise l'expression « habileté » plutôt que l'expression « trait ». On considérera donc, dans cette étude, que ce sont deux expressions conceptuellement équivalentes.

2. C'est-à-dire, celle qui a cours à une époque donnée et dans un milieu scientifique donné.

3. Pour la dérivation des expressions exactes de σ_{r_i} et de $\sigma^2_{U'_k}$, on peut consulter Stout (1987), Nandakumar & Stout (1993), Blais & Laurier (1995).

4. Environ 5 % des participants présentaient des schémas de réponses aberrants. Ils ont été repérés en utilisant un indicateur de reproductibilité obtenu grâce à la création d'une hiérarchie de Guttman. Ils ont par la suite été exclus des analyses.

5. Dans le premier sous-test, on retrouve six items qui n'ont pu être classés dans l'un ou l'autre des domaines.

6. C'est-à-dire une procédure pour laquelle existe un programme informatique disponible commercialement et dont la manipulation ne demande pas des connaissances très pointues en informatique.

RÉFÉRENCES

Bejar, I.I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameters estimates. Journal of Educational Measurement, 17, 283-296.

Blais, J.-G. (1987). Effets de la violation du postulat d'unidimensionalité dans la théorie des réponses aux items. Thèse de doctorat non publiée, Université de Montréal.

Blais, J.-G. & Laurier, M.D. (1997). Content considerations and resampling strategies in assessing unidimensionality of set of items. Actes du colloque 1995 de l'International Association for Educational Measurement, 361-370.

Blais, J.-G. & Laurier, M.D. (1995). Methodological considerations in using DIMTEST to assess unidimensionality. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, San Francisco.

Bock, R.D., Gibbons, R.D. & Muraki, E. (1985). Full information factor analysis, MRC Report 85-1. Chicago, Il : Methodology Research Center, National Opinion Research Center.

Chen, W.-H. & Thissen, D. (1997). Local dependance indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics, 22, 265-289.

Cliff, N. (1983). Evaluating Guttman scales : Some old and new thoughts. In H. Wainer & S. Messick (réd.), Principles of modern psychological measurement (pp. 283-301). Hillsdale, NJ : Lawrence Erlbaum.

de Champlain, A. & Gessaroli, M.E. (1991). Assessing test dimensionality using an index based on nonlinear factor analysis. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Chicago.

Doody-Bogan, E. & Yen, W.M. (1983). Detecting multidimensionality and examining its effect on vertical equating with the three parameter logistic model. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Montréal.

Drasgow, F. & Parsons, C.K. (1983). Application of unidimensional psychological item response theory models to multidimensional data. Applied Psychological Measurement, 7, 189-199.

Goldstein, H. (1980). Dimensionality, bias, independance and measurement scale problems in latent trait test score models. British Journal of Mathematical and Statistical Psychology, 33, 234-246.

Greaud, V.A. (1988). Some effects of applying unidimensional IRT to multidimensional tests. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Nouvelle-Orléans.

Hambleton, R.K. (1989). Principles and selected applications of IRT. In R.L. Linn (réd.), Educational Measurement (pp. 147-200). New York : American Council on Education/MacMillan.

Hambleton, R.K. & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10, 287-302.

Harrison, D.A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. Journal of Educational Statistics, 11, 91-115.

Hattie, J.A. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

Hattie, J.A. (1985). Methodology review : assessing unidimensionality of test and items. Applied Psychological Measurement, 9, 139-164.

Hattie, J.A., Krakowski, K., Rogers, H.J. & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. Applied Psychological Measurement, 20, 1-14.

Henning, G.T., Hudson, T. & Turner, J. (1985). Item response theory and the assumption of unidimensionality. Language Testing, 2, 141-154.

Holland, P.W. & Rozenbaum, P.R. (1986). Conditional association and unidimensionality in monotone talent trait variable models. Annals of Statistics, 14, 1523-1543.

Humphreys, L. (1984). A theoretical and empirical study of the psychometric assessment of psychological test dimensionality and bias (ONR Research Proposal). Washington, DC : Office of Naval Research.

Jöreskog, K.W. & Madsen, M.S. (1983). LISREL user's guide (logiciel). Uppsala : Département de statistique, Université d'Uppsala, Suède.

Kim, H.R. & Stout, W.F. (1993). A robustness study of ability estimation in the presence of latent trait multidimensionality using the Junker/Stout index of dimensionality. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Atlanta.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, N.J. : Lawrence Erlbaum Associates.

Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, MA : Addison-Wesley.

Luecht, R.M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. Applied Psychological Measurement, 20, 389-404.

McDonald, R.P. (1981). The dimensionality of test and items. British Journal of Mathematical and Statistical Psychology, 27, 82-89.

McDonald, R.P. (1989). Future directions for item response theory. International Journal of Educational Research,13, 205-220.

McNemar, Q. (1946). Opinion-attitude methodology. Psychological Bulletin, 43, 289-374.

Messick, S. (1989). Validity. In R.L. Linn (éd.), Educational Measurement (3e éd.) (pp. 13-103). New York : Macmillan.

Mislevy, R.J. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.

Nandakumar, R. & Stout, W. (1993). Refinement of Stout's procedure for assessing latent trait unidimensionality. Journal of Educational Statistics, 18, 41-68.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests : results and implications. Journal of Educational Statistics, 4, 207-230.

Reckase, M.D. (1990). Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Boston.

Roussos, L.A., Stout, W. & Marden, J.I. (1993). Analysis of the multidimensional structure of standardized test using DIMTEST with hierarchical cluster analysis. Manuscrit non publié, Université de l'Illinois, Département de statistique.

Rozenbaum, P.R. (1984). Testing the conditional independance and monotonicity assumptions of item response theory. Psychometrika, 49, 425-436.

Stout, W. (1987). A non-parametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. Psychometrika, 55, 293-326.

Stout, W., Nandakumar, R., Junker, B. & Chang, H.H. (1991). DIMTEST and TESTSIM (logiciels). Champaign : Département de statistique, Université de l'Illinois.

van der Linden, W.J. (1996). Assembling tests for the measurement of multiple traits. Applied Psychological Measurement, 20, 373-388.

Wang, M. (1988). Measurement bias in the application of a unidimensional model to multidimensional item-response data. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Nouvelle-Orléans.

Wilson, D., Wood, R.L. & Gibbons, R.D. (1991). Testfact : test scoring and item factor analysis (logiciel). Chicago, Il : Scientific Software.

Zuroff, D.C. (1986). Was Gordon Allport a trait theorist? Journal of Personality and Social Psychology, 51, 993-1000.

Zwick, R. (1987). Assessment of the dimensionality of year 15 reading data. In A.E. Beaton (réd.), Implementing the new design : The NAEP 1983-1984 Technical report (pp. 245-284). Rapport 25-TR-20. Princeton, NJ : National Assessment of Educational Progress.