Etude de la dimensionnalité d'un test de raisonnement à l'aide des Modèles de Réponse à l'Item

Jacques Juhel (*)

Summary : *Several Item Response Theory (IRT) based methods can be used for studying the dimensionality of dichotomous data. One such typically selected approach is the verification of the unidimensionality of a set of items with unidimensional IRT. The identification of data's multidimensional structure with nonlinear factorial analysis or multidimensional IRT is another promising research strategy. Some of these models are first briefly presented. Their potential in studying the dimensionality of a set of dichotomous data collected with items of a reasoning test is then illustrated. Results of data analysis give evidence in support of multidimensional IRT in assessing dimensionality of test data and in the measurement of individual differences.*

Key words : *Dimensionality, nonlinear factorial analysis, multidimensional item response theory (MIRT), confirmatory research strategy*

Résumé : *Plusieurs méthodes basées sur les Modèles de Réponse à l'Item (MRI) peuvent être employées pour étudier la dimensionnalité de données dichotomiques. La stratégie peut consister soit à vérifier l'unidimensionnalité de données, soit à identifier la structure multidimensionnelle des données grâce à des modèles comme l'analyse factorielle non linéaire ou les MRI multidimensionnels. Après avoir brièvement présenté certains de ces modèles, on tente ensuite d'en illustrer le potentiel en les appliquant à des données recueillies à l'aide d'un test de raisonnement comportant trois échelles d'items dichotomiques. Les ré-*

* Université de Rennes 2 – CRPCC, Laboratoire de psychologie expérimentale, Groupe de recherche en psychologie différentielle – 6, avenue G. Berger – 35043 Rennes Cedex (Mel: jacques.juhel@uhb.fr).

sultats des analyses effectuées témoignent des possibilités de modélisation offertes et illustrent l'intérêt des MRI dans l'étude de tests multidimensionnels et dans la mesure des caractéristiques individuelles.

Mots Clés : *Dimensionnalité, analyse factorielle non linéaire, modèles multidimensionnels de réponse à l'item, stratégie de recherche confirmatoire*

INTRODUCTION

Les Modèles de Réponse à l'Item (MRI) dont l'objectif général est de rendre compte des réponses d'individus à un ensemble donné d'items traitent fondamentalement de la relation entre la probabilité $P(X_{ij}=1|a_j,b_j,c_j,\theta_i)$ d'observer une réponse correcte de l'individu *i* à l'item *j* et certaines caractéristiques de l'individu et de l'item (Rasch, 1960/80). Ces modèles mathématiques, élaborés originellement pour des données dichotomiques sous-tendues par une seule dimension, représentent le plus souvent la *fonction de réponse à l'item* sous la forme d'une régression non linéaire de l'item sur le trait latent mesuré[1] (Hambleton, 1989 ; Hambleton et Swaminathan, 1985 ; Lord, 1980 ; Lord et Novick, 1968 ; van der Linden et Hambleton, 1997 ; Wright et Stone, 1979). La fonction de réponse à l'item d'un MRI paramétrique appliqué à des données dichotomiques est ainsi classiquement définie comme la probabilité d'observer une réponse correcte de l'individu *i* à l'item *j* en fonction de plusieurs paramètres de l'item *j* (la puissance discriminative a_j, la difficulté b_j, l'asymptote basse ou le paramètre de pseudo-chance c_j) et de l'individu *i* (le niveau d'aptitude θ_i). Cette fonction est plus ou moins complexe selon le type de relation spécifiée[2] et le nombre de paramètres considérés (voir pour une présentation en français Dickes, Flieller, Tournois et Kop, 1994).

Comme pour tous les modèles qui spécifient une relation entre des observations et une ou plusieurs constructions psychologiques, l'application des MRI à un ensemble de données nécessite que certains postulats soient respectés. Ces exigences sont bien sûr moins grandes dans le cas de modèles non paramétriques. Tous les MRI reposent ainsi sur l'hypothèse de monotonie, hypothèse selon la-

[1] Quelques MRI dits non paramétriques n'attribuent cependant aucune forme particulière à la fonction de réponse à l'item (Sijtsma, 1998).

[2] La fonction de répartition généralement choisie a une densité de probabilité normale (fonction ogive normale) ou logistique (fonction logistique).

quelle la probabilité de réussite à l'item croît avec le niveau d'aptitude. Beaucoup de ces modèles font également l'hypothèse d'unidimensionnalité des données. L'hypothèse d'une seule aptitude est ainsi jugée suffisante pour rendre compte des réponses observées à l'ensemble d'items considéré. Enfin, l'hypothèse d'indépendance locale explique les relations entre items par leur seule dépendance au trait latent, l'aptitude de l'individu et les caractéristiques de l'item étant alors considérées comme les seuls facteurs de la performance.

Bien que liées, les hypothèses d'indépendance locale et d'unidimensionnalité ne sont pas assimilables l'une à l'autre et la première est utilisable aussi bien dans les MRI unidimensionnels que dans ceux qui ne font pas l'hypothèse d'unidimensionnalité (MRI multidimensionnels). La dimensionnalité *k* d'un ensemble d'items est en effet définie dans les MRI comme le nombre total de traits latents nécessaires et suffisants à l'explication des relations entre items[3]. Ces *k* traits latents constituent un espace *k*-dimensionnel au sein duquel il est possible de situer l'individu au moyen du vecteur de scores $(\theta_1,\theta_2,...,\theta_k)$. L'indépendance locale est généralement considérée dans les MRI unidimensionnels comme un cas spécial d'unidimensionnalité de l'espace latent (Hambleton et Swaminathan, 1985 ; Lord et Novick, 1968). Mais les items d'un MRI explicitement unidimensionnel peuvent ne pas être localement indépendants. C'est par exemple le cas lorsque la probabilité de bonne réponse à un item donné dépend du patron des réponses fournies aux items antérieurement présentés[4] ou lorsqu'elle dépend de dimensions mineures, spécifiques à un nombre réduit d'items. Un MRI localement indépendant peut aussi être multidimensionnel au sens où plusieurs composantes du trait latent (e.g., des processus dans un modèle composantiel unidimensionnel) ou plusieurs traits latents, éventuellement interdépendants, peuvent influencer les réponses observées. L'hypothèse d'unidimensionnalité est donc fonction du niveau d'approximation choisi.

Stout (1987), en suggérant de distinguer une forme " essentielle " d'indépendance, a proposé une définition moins stricte de l'indépendance locale

[3] Dans le cas de données dichotomiques, la dimensionnalité peut être techniquement définie de la manière suivante (Nandakumar, 1991). Soient *X*ij la réponse (1 pour réussite, 0 pour échec) de l'individu *i* à l'item *j* et *X*n le vecteur de réponses à un ensemble donné de *n* items. Si *k* traits latents interviennent dans un MRI localement indépendant et s'il est impossible de produire un tel modèle localement indépendant pour *X*n avec moins de *k* traits latents, la dimensionnalité de *X*n est *k*. Le MRI est dit unidimensionnel si *k*=1, multidimensionnel si *k*>1.Cette définition de la dimensionnalité impose donc que toutes les covariances entre items soient nulles lorsque les *k* traits latents sont maintenus constants (covariances conditionnelles aux *k* traits latents).

[4] Voir à ce sujet les travaux sur les MRI dynamiques (e.g., Verhelst et Glas, 1993) ou ceux sur les modèles à dépendance locale (e.g., Hoskens et De Boeck, 1997).

qui s'inscrit dans le contexte des MRI non paramétriques (Mokken et Lewis, 1982). Pour Stout, un ensemble de données est essentiellement dépendant par rapport à k_E traits latents si pour un ensemble donné de réponses à n items, la valeur moyenne des valeurs absolues des covariances conditionnelles aux k_E traits latents entre tous les items tend vers 0 quand n tend vers l'infini. Contrairement donc à la définition de la dimensionnalité dans les MRI paramétriques, cette définition plus libérale met l'accent sur la seule influence des traits " dominants " en considérant que l'hypothèse d'unidimensionnalité essentielle ($k_E = 1$) est respectée si une dimension dominante, bien distincte de dimensions mineures, peut être identifiée dans les données (Nandakumar, 1991, 1993).

La dimensionnalité d'un test ou d'un ensemble d'items est souvent difficile à juger. Il est bien sûr toujours possible en théorie de sélectionner des items afin de mesurer une seule dimension. Mais le tableau est rarement aussi simple en pratique car la dimensionnalité d'un modèle ou d'un ensemble de données repose fondamentalement sur l'influence conjointe des items et des individus qui y répondent. Certains facteurs individuels comme les choix procéduraux, la gestion des échanges entre vitesse et précision, le niveau d'anxiété et de motivation, etc. peuvent influencer les réponses aux items. Une certaine hétérogénéité peut aussi se retrouver au niveau du contenu et de la nature des items, certains items étant par exemple plus sensibles que d'autres aux différences individuelles dans le niveau d'aptitude. Les items et les individus produisent donc ensemble des données dont la dimensionnalité doit être systématiquement interrogée (e.g., Hambleton et Rovinelli, 1986 ; Hambleton, Swaminathan et Rogers, 1991 ; voir aussi les articles de McDonald, Reckase et Gessaroli dans l'ouvrage dirigé par Laveault, Zumbo, Gessaroli et Boss, 1994). Cette interrogation peut s'appuyer sur diverses méthodes.

METHODES D'EVALUATION DE LA DIMENSIONNALITE DE DONNEES DICHOTOMIQUES

L'unidimensionnalité d'un test ou de chacun des sous-tests qui le composent peut être vérifiée à l'aide des MRI unidimensionnels. L'emploi de ceux-ci doit néanmoins être approximativement justifié en appliquant préalablement par exemple une procédure non paramétrique d'évaluation de la dimensionnalité des items considérés. Une seconde étape doit succéder à la précédente si les analyses effectuées conduisent à réfuter l'hypothèse d'unidimensionnalité. L'identification de la structure multidimensionnelle du test impose alors de faire appel à d'autres méthodes comme l'analyse factorielle linéaire des corrélations inter-items ou mieux, l'analyse factorielle non linéaire et les MRI multidimensionnels.

L'utilisation conjointe de ces méthodes, associées ou non à certaines procédures complémentaires (e.g., analyse en clusters, échelonnement multidimensionnel, etc.), peut aussi apporter des informations utiles à la compréhension de la dimensionnalité des données.

Les procédures non paramétriques

L'utilisation d'approches non paramétriques ne faisant l'hypothèse d'aucune fonction paramétrique de réponse à l'item et dans lesquelles la dimensionnalité est définie de manière moins contraignante que dans les MRI paramétriques peut être une étape utile dans la détermination de la dimensionnalité d'un ensemble d'items. Ces diverses procédures sont basées sur un principe d'estimation de covariances inter-items conditionnelles au niveau d'aptitude de l'individu. Elles peuvent être employées pour évaluer le manque d'unidimensionnalité d'un ensemble d'items dichotomiques, identifier des sous-groupes d'items essentiellement unidimensionnels ou estimer la " quantité " de multidimensionnalité présente dans les données (pour revue, Stout, Douglas, Kim, Roussos et Zhang, 1996).

La procédure DIMTEST (Nandakumar et Stout, 1993 ; Stout, 1987) dont nous présenterons une application plus loin en est un exemple. Elle consiste à partager l'ensemble des items considérés en plusieurs sous-tests. Les items du premier sous-test (AT1) sont choisis (par l'utilisateur ou automatiquement après analyse en facteurs communs des corrélations tétrachoriques ou classification hiérarchique) comme mesurant le même trait dominant. Les items du second sous-test (AT2), en nombre égal à ceux de AT1, sont ensuite sélectionnés de telle manière que leur difficulté soit distribuée de la même façon que celle des items de AT1. Les items restants constituent le sous-test de partitionnement (PT). Ils servent à répartir les sujets dans différents sous-groupes en fonction de leur score au test. Une première statistique (T_1) est calculée pour les items de AT1. Celle-ci peut être comprise comme la moyenne sur l'ensemble des sous-groupes de sujets de la différence calculée au sein de chaque sous-groupe, entre deux estimations de la variance. La première variance est celle des scores observés. La seconde est une estimation de la variance des scores sous hypothèse d'unidimensionnalité. Elle est égale à la première variance si l'hypothèse d'unidimensionnalité est respectée et lui est inférieure dans le cas contraire. Une procédure semblable permet d'obtenir une seconde statistique (T_2) pour les items de AT2. L'hypothèse d'unidimensionnalité peut alors être testée à partir de la statistique $T = T_1 - T_2$, de distribution asymptotiquement normale, de moyenne nulle et de variance 1. L'idée est que si le modèle sous-tendant les réponses aux items est essentiellement unidimensionnel, l'écart entre T_1 et T_2 doit être proche de 0. Si par contre, le modèle qui sous-tend

les réponses au test n'est pas essentiellement unidimensionnel, les items de AT1 sont dimensionnellement différents de ceux de AT2 et de PT, la valeur du T de Stout est importante et la probabilité de rejet de l'hypothèse nulle, forte. Signalons par ailleurs que cette procédure semble être peu sensible à l'existence de corrélations entre traits latents et à la méthode d'affectation des items dans les différents sous-tests (Hattie, Krakowski, Rogers et Swaminathan, 1996).

Les MRI unidimensionnels

Un MRI unidimensionnel n'est généralement appliqué à des données que si l'on a quelque raison de penser qu'une seule dimension en sous-tend l'organisation. Qu'il s'agisse du modèle de Rasch ou des modèles logistiques à 2 ou à 3 paramètres, l'hypothèse d'unidimensionnalité est considérée comme satisfaite si le modèle s'ajuste aux données c'est-à-dire si l'écart entre les valeurs observées et les valeurs prédites par le modèle pour les items et les individus est faible. L'évaluation de l'unidimensionnalité, synonyme ici d'évaluation de l'adéquation du modèle, repose donc sur l'estimation préalable des paramètres du MRI unidimensionnel choisi. Ces paramètres peuvent être ensuite comparés entre différents sous-groupes de sujets ou d'items selon que l'on étudie l'invariance des paramètres de l'item ou de l'individu. L'égalité des paramètres estimés est testée au moyen de statistiques obtenues en comparant les distributions des scores observés et des scores prédits sous hypothèse d'ajustement du MRI. De nombreuses mesures d'ajustement globales (modèle) ou analytiques (items, individus), souvent basées sur le chi-deux ou le t, ont été proposées (pour revue Flieller, 1993 ; Hattie, 1985). Ces indices d'ajustement semblent néanmoins rarement satisfaisants en eux-mêmes (distribution d'échantillonnage souvent mal connue, sensibilité à la taille de l'échantillon, critère de décision rarement fondé).

Il est donc souhaitable de ne pas évaluer l'ajustement d'un MRI unidimensionnel, et donc l'hypothèse d'unidimensionnalité, à l'aide de ces seuls indices. Hambleton, Swaminathan et Rogers (1991) proposent par exemple de compléter l'information fournie par les indices d'ajustement en combinant plusieurs approches comme l'analyse factorielle d'items, l'analyse des covariances entre items au sein de groupes de niveau d'aptitude homogène, l'étude de l'homogénéité des corrélations point-bisériales ou l'examen comparatif des résidus pour les MRI unidimensionnels à 1, 2 et 3 paramètres.

L'analyse factorielle linéaire des corrélations entre items

L'identification des facteurs susceptibles de sous-tendre l'organisation de données dichotomiques au moyen de l'analyse factorielle linéaire n'est pas sans

poser de sérieuses difficultés. Lord et Novick (1968) signalent ainsi qu'il est contradictoire de définir une variable aléatoire discrète comme une combinaison linéaire de variables latentes continues. On sait en outre que l'analyse factorielle linéaire de coefficients phi conduit fréquemment à une surestimation de la dimensionnalité en raison de l'émergence de dimensions artefactuelles[5] (e.g., la " difficulté " des items du test) s'expliquant par la non linéarité des fonctions de régression des items sur le trait latent (Bernstein et Teng, 1989 ; Hulin *et al.* 1983, cités dans Nandakumar et Stout, 1993).

Une approche parfois recommandée consiste à évaluer approximativement la dimensionnalité des données à partir de l'analyse des corrélations tétrachoriques[6] entre items (Reckase, 1979 ; Lord, 1980). Si la première valeur propre est élevée par rapport à la seconde et que celle-ci n'est pas plus importante que les valeurs propres de rang supérieur, il est alors possible de conclure à l'unidimensionnalité des données. Une façon plus fine de procéder repose sur la comparaison de ces valeurs propres à celles d'une matrice de corrélations de données aléatoires obtenues avec un nombre identique de variables sur un échantillon de même taille (Hambleton, Swaminathan et Rogers, 1991). L'hypothèse d'unidimensionnalité peut être retenue lorsque, à l'exception de la première d'entre-elles, les valeurs propres sont les mêmes. Il semble cependant que les corrélations tétrachoriques entre items peuvent être biaisées quand il est possible de répondre correctement au hasard (Carroll, 1945), quand la distribution du trait latent n'est pas normale ou quand la fonction de réponse à l'item n'est pas la fonction ogive normale (Lord, 1980).

Les approches multidimensionnelles

La modélisation de données dichotomiques à l'aide de MRI unidimensionnels localement indépendants n'est bien sûr pas toujours possible. On peut alors, lorsque l'exigence d'unidimensionnalité s'avère - trop - incompatible avec l'organisation des données, faire appel à des approches paramétriques multidimensionnelles pour déterminer la dimensionnalité de matrices de données binaires. Deux catégories de méthodes dont les formulations statistiques sont virtuellement identiques peuvent être utilisées à cette fin (Reckase, 1997b). Les méthodes

[5] Un facteur est dit artefactuel (*spurious*) quand il provient des propriétés de mesure des variables observées et de la distribution des réponses à l'item sur les différentes catégories de réponse plutôt que de la structure latente des données.

[6] La corrélation tétrachorique entre deux items j_1 et j_2 est une estimation, calculée par exemple par la méthode du maximum de vraisemblance, de la corrélation entre les variables latentes continues X_{j1} et X_{j2} qui sous-tendent j_1 et j_2.

d'analyse factorielle non linéaires sont généralement employées pour identifier les dimensions qui résument le mieux les relations entre items dichotomiques. Elles peuvent être utilisées aussi bien dans une logique exploratoire que pour évaluer l'adéquation aux données de représentations factorielles élaborées *a priori* (logique confirmatoire, Dickes, 1996). Les MRI multidimensionnels présentent l'avantage par rapport à l'analyse factorielle d'items de plus se préoccuper de la description des caractéristiques des items et de leur interaction avec les traits latents mesurés.

Les méthodes d'analyse factorielle non linéaire

Contrairement au modèle d'analyse factorielle classique pour variables continues qui fait l'hypothèse d'un processus de réponse directement observable, les modèles d'analyse factorielle non linéaire pour variables dichotomiques font l'hypothèse d'un processus de réponse non observable y_{ij} , pour l'individu i et l'item j, défini comme une combinaison linéaire de m variables latentes θ_{ki} de distribution normale, pondérées par les saturations λ_{jk} soit :

$$y_{ij} = \lambda_{j1}\theta_{1i} + \lambda_{j2}\theta_{2i} + ... + \lambda_{jm}\theta_{mi} + \delta_i$$

Ces modèles postulent l'existence d'une variable continue non observable y_{ij} , dichotomisée en un score observé 1 ou 0 selon que le niveau de compétence du sujet est inférieur ou supérieur à un certain seuil γ_j pour l'item j soit :

$$x_{ij} = \begin{cases} 1 \text{ si } y_{ij} \geq \gamma_j \\ 0 \text{ si } y_{ij} < \gamma_j \end{cases}$$

Deux modèles d'analyse factorielle non linéaire principaux ont été développés. On peut les différencier selon que l'exigence d'indépendance locale est faible ou forte[7] (*weak* vs *strong local independence*) c'est-à-dire selon la nature de l'information analysée (*limited* vs *full information*).

Le premier de ces modèles a été proposé par McDonald (1967, 1982, 1997). C'est un modèle à ogive normale $N(.)$ accompagné d'un principe d'indépendance locale faible, l'information analysée étant celle de la matrice de moment-produit[8] de l'échantillon. Le modèle définit la probabilité de bonne réponse à l'item j par l'individu i par l'équation :

[7] L'exigence d'indépendance locale est forte (resp. faible) si les k composantes de θ rendent théoriquement compte de toutes les relations entre les *probabilités de bonne réponse* (resp. de toutes les *covariances* entre items). Dans ce second cas, les covariances résiduelles entre toutes les paires d'items à des niveaux d'aptitude fixés sont toutes nulles lorsque l'espace latent est unidimensionnel.

[8] La matrice de moment-produit est obtenue en multipliant la matrice de données binaires par sa transposée.

$$P(x_{ij} = 1|\theta_i) = N(\beta_{j0} + \beta_{j1}\theta_{1i} + ... + \beta_{jk}\theta_{ki})$$

avec β_{jk} élément de la matrice de structure $B = [\beta_{jk}]$ et *k* facteurs standardisés θ_{ki}. Ce modèle, qui peut être représenté sous forme d'une régression polynomiale infinie des données sur les facteurs[9], peut être approximé par la méthode des moindres carrés. McDonald a montré qu'il est possible de déduire les paramètres de seuil[10] γ_j (interprétables en termes de difficulté) et les saturations λ_{jk} des items à partir des estimations des paramètres β_{jk} et des covariances entre facteurs. Ces paramètres peuvent être estimés avec le programme NOHARM (*Normal Ogive Harmonic Analysis Robust Method*) développé par Fraser et McDonald (1988) dont la méthode d'estimation est celle des moindres carrés non pondérés (ULS : *Unweighted Least Squares*). C'est une méthode robuste, peu sensible au non respect de l'hypothèse de normalité du vecteur-trait latent mais qui ne fournit ni estimation des erreurs-type de mesure, ni test statistique permettant de juger de l'ajustement du modèle. La détermination de la dimensionnalité des données repose alors sur l'examen comparatif des covariances résiduelles entre items pour chacune des solutions factorielles testées. Cet examen peut être facilité par l'utilisation d'une statistique du chi-deux qui permet de tester l'hypothèse H_0 que les éléments de la matrice résiduelle sous-diagonale sont nuls (De Champlain et Linda Tang, 1997).

Une seconde méthode d'analyse factorielle plus directement basée sur les MRI a été développée par Bock et ses collaborateurs (Bock et Aitkin, 1981 ; Bock, Gibbons et Muraki, 1988). Cette extension du modèle à ogive normale à 2 paramètres, connue sous le nom de *Full Information item Factor Analysis*, peut être vue comme un cas plus général de l'analyse factorielle (Dickes, ce numéro ; Reckase, 1997b) dans la mesure où on cherche à reproduire toute l'information de la matrice de données binaires (exigence d'indépendance locale forte). La probabilité de bonne réponse du sujet *i* à l'item *j* y est définie par l'équation:

$$P(x_{ij} = 1|\theta_i) = \Phi_j(\theta_{ki}),$$

$\Phi_j(.)$ étant une fonction ogive normale des paramètres de discrimination a_{jk} (ou pente) et de difficulté b_j (ou ordonnée à l'origine) de l'item *j* à partir desquels peuvent être calculés les saturations λ_{jk} et les seuils γ_j. Un troisième paramètre c_j

[9] La fonction ogive normale dont la courbe représentative est très peu différente de celle de la fonction logistique permet le développement de séries harmoniques du type :

$P(x_{ij} = 1|\theta_i) = f_{j0} + f_{j1}\theta_i + f_{j2}\theta_i^2 + ...$ où f_{jk} est la saturation du facteur *f* sur l'item *j*.

[10] Ces paramètres de seuil sont les écarts normaux spécifiant l'aire sous la courbe normale égale au pourcentage de réponses incorrectes aux items.

(asymptote basse) peut également être ajouté au modèle. Un programme (TESTFACT - Bock, Gibbons et Muraki, 1988; Wilson, Wood et Gibbons, 1991) permet l'estimation des paramètres de l'item par la méthode du maximum de vraisemblance marginal (MML : *marginal maximum likelihood*). La dimensionnalité des données peut être explorée en testant successivement des modèles de dimensionnalité croissante et en évaluant la contribution de chaque facteur supplémentaire par le gain d'ajustement apporté et mesuré par un χ^2 partiel du rapport de vraisemblance. L'analyse se poursuit jusqu'à ce que l'introduction d'un facteur supplémentaire ne produise plus d'amélioration significative de l'ajustement du modèle.

Les MRI multidimensionnels

Les MRI multidimensionnels pour items dichotomiques sont des extensions des MRI unidimensionnels à 1, 2 et 3 paramètres. Comme ces derniers, ils visent à fournir une description adéquate de l'interaction individu-item (Lord, 1980) et une estimation multidimensionnelle des caractéristiques des items et des individus en remplaçant l'hypothèse d'unidimensionnalité par celle, moins contraignante, d'adéquation entre la dimensionnalité du MRI et la dimensionnalité du test (McKinley, 1988 ; McKinley et Reckase, 1983 ; Reckase, 1979, 1997a,b). On trouvera dans les ouvrages collectifs édités par Engelhard et Wilson (1996), van der Linden et Hambleton (1997), Wilson (1992, 1994) ou Wilson, Engelhard et Draney (1997) de nombreux exemples d'application de ces modèles à l'analyse de tests multidimensionnels, à l'analyse composantielle d'épreuves cognitives, à la modélisation de processus de changement et d'apprentissage, etc.

Ackerman (1996) propose de distinguer deux grandes catégories de modèles selon qu'ils appliquent ou non un principe de compensation entre dimensions (ou composantes). Les modèles dits compensatoires font l'hypothèse d'autant de paramètres de discrimination que de dimensions mais d'un seul paramètre de difficulté de l'item. Un score élevé par rapport à une dimension peut donc compenser un score faible par rapport à une autre. Cette hypothèse n'est pas faite dans les modèles non compensatoires dont les termes sont multiplicatifs. Un paramètre de discrimination et un paramètre de difficulté sont alors associés à chaque dimension. Mathématiquement, les MRI multidimensionnels compensatoires dont nous donnerons un exemple d'application plus loin utilisent la distribution logistique. Le parallèle avec l'analyse factorielle est donc moins immédiat qu'il ne l'est avec les modèles d'analyse factorielle non linéaire. L'équation du modèle définit une surface de réponse à l'item qui donne la probabilité de réponse correcte à l'item en fonction de la localisation des sujets dans l'espace des aptitudes

spécifié par le vecteur θ_i. Reckase (1997a) en propose par exemple la formulation suivante :

$$P\left(X_{ij}=1\middle|\theta_i\right)=c_j+(1-c_j)\frac{e^{(a_j'\theta_i+d_j)}}{1+e^{(a_j'\theta_i+d_j)}}$$

où a_j' est le vecteur des paramètres de discrimination de l'item, d_j le paramètre de difficulté de l'item et θ_i le vecteur d'aptitudes de l'individu *i*. L'introduction dans l'exposant du modèle d'une matrice de structure de l'item S_j identifiant les dimensions mesurées permet son utilisation dans une logique confirmatoire[11] (McKinley, 1989).

L'estimation des paramètres de l'item pour des modèles de ce type peut être effectuée par exemple avec TESTMAP (McKinley, 1992) qui utilise la méthode MML. La procédure d'estimation est adaptée de celle utilisée dans le cas unidimensionnel (procédure EM de BILOG 3). Le programme fournit aussi des indices d'entropie sur lesquels l'utilisateur peut s'appuyer pour évaluer l'adéquation du modèle (AIC : *Akaike Information Criterion* ; CAIC : *Consistent AIC*). La stratégie confirmatoire recommandée pour tester des hypothèses de dimensionnalité consiste alors à comparer différents modèles du point de vue de leur adéquation aux données à partir d'une correspondance établie *a priori* entre les dimensions du modèle et les dimensions du test, identifiées sur la base d'une analyse des contenus (ou mieux, des processus) impliqués dans les items.

ILLUSTRATION

Nous appliquerons maintenant certaines des méthodes qui viennent d'être brièvement présentées à l'étude de la dimensionnalité de données dichotomiques recueillies à l'aide d'un test de raisonnement comportant 45 items à choix multiple administrés à 8685 jeunes adultes lors d'un concours de niveau baccalauréat. Les items de ce test sont répartis en trois échelles (voir figure 1 pour des exemples d'items).

La première échelle de " Logique Opératoire " (LO) comporte 16 items dont la résolution nécessite d'appliquer une ou plusieurs transformations à un symbole donné en fonction de règles à découvrir à partir d'informations apparaissant dans un tableau. La seconde échelle de " Raisonnement Spatial " (RS) com-

[11] L'expression $a_j'\theta_i+d_j$ devient alors $a_j'S_j\theta_i+d_j$ où S_j est une matrice $k\times k$ spécifiant la ou les dimensions mesurées par l'item.

prend 15 items. La tâche consiste ici à faire mentalement la synthèse de grilles de 9 cases afin de trouver celles qui, superposées les unes aux autres, forment la grille cible ; 3 grilles sont à identifier parmi les 6 présentées. La troisième échelle de " Transformation de Lettres " (TL) est composée de 14 items. Pour chaque item, il s'agit de compléter une série de lettres et d'appliquer au résultat provisoirement obtenu une règle de transformation donnée.

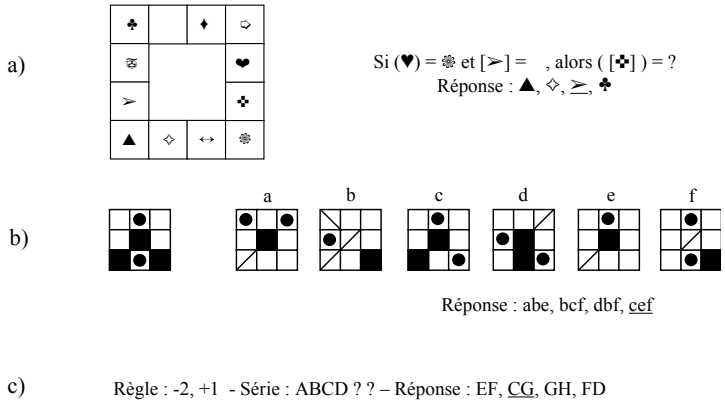

Figure 1
Exemples d'items des échelles de Logique Opératoire (a), Raisonnement Spatial (b) et Transformation de Lettres (c).

Ce test de raisonnement étant par construction, constitué de trois échelles, il est théoriquement et pratiquement important de s'assurer de l'unidimensionnalité de chacune d'entre-elles. Plusieurs séries d'analyses conduites avec cet objectif sont décrites par la suite. L'hypothèse d'unidimensionnalité " essentielle " est d'abord testée avant d'appliquer les MRI unidimensionnels puis deux modèles multidimensionnels. Les résultats présentés servent essentiellement ici à illustrer l'intérêt des MRI dans l'étude de la dimensionnalité d'un ensemble de réponses dichotomiques réellement observées.

Test de l'hypothèse d'unidimensionnalité " essentielle "

L'existence d'une dimension dominante par échelle est étudiée à l'aide de la procédure DIMTEST présentée précédemment. Les données dichotomiques auxquelles on a appliqué le programme DIMTEST (Stout, Douglas, Junker et Roussos, 1993) proviennent d'un sous-échantillon aléatoire de 2607 sujets. L'objectif étant d'évaluer dans quelle mesure les items de Logique Opératoire, de raisonnement Spatial et de Transformation de Lettres sont dimensionnellement différents les uns des autres, trois analyses ont été effectuées en affectant à chaque

fois tous les items d'une même échelle (LO, TL, RS) au sous-test d'évaluation AT1, le sous-test de partitionnement étant alors composé d'items des deux autres échelles.

Tableau 1
Statistiques fournies par le programme DIMTEST (LO : 16 items de Logique Opératoire ; RS : 15 items de Raisonnement Spatial ; TL : 14 items de Transformation de Lettres ; N = 2607).

Items composant le sous-test AT1	Items composant les sous-tests AT2 et PT	T_1	T_2	*T de Stout*	*p*
LO	TL+RS	13,842	9,375	4,467	.00000
RS	TL+LO	11,917	5,681	6,236	.00000
TL	LO+RS	21,987	10,623	11,365	.00000

Comme on peut le voir dans le tableau 1, les valeurs du T de Stout sont toutes significativement différentes de 0. Ces résultats conduisent donc comme attendu à rejeter l'hypothèse d'unidimensionnalité de l'ensemble des 45 items. Ils montrent parallèlement qu'en comparaison à un ensemble d'items dont la difficulté est distribuée semblablement, les échelles de Logique Opératoire, de Raisonnement Spatial et de Transformation de Lettres sont respectivement essentiellement unidimensionnelles. Cette première analyse suggère donc que chacune des trois échelles mesure une dimension dominante sans exclure pour autant l'existence de dimensions mineures pouvant sous-tendre les réponses aux items du test de raisonnement.

Vérification de l'hypothèse d'unidimensionnalité avec les MRI unidimensionnels

On pourrait souhaiter prolonger les résultats précédents, obtenus à un niveau d'analyse inter-échelles, en appliquant DIMTEST aux seuls items de chacune des 3 échelles. Il suffirait alors d'affecter un certain nombre d'items d'une même échelle à AT1 et d'appliquer la procédure autant de fois que l'on modifie le contenu de AT1 (items pairs, impairs, faciles, difficiles, etc.). Nandakumar (1993) recommande cependant de n'utiliser la procédure DIMTEST qu'avec 25 items au moins car la précision du T de Stout semble être très incertaine lorsque cette condition n'est pas remplie.

La procédure précédente ne pouvant être appliquée sur ces données, on peut utiliser la stratégie qui consiste à estimer *sous hypothèse d'unidimensionnalité* les paramètres d'un MRI unidimensionnel pour vérifier ensuite la validité des postulats du modèle utilisé. Mais nous avons dit que la conception de

l'unidimensionnalité classiquement énoncée dans les MRI unidimensionnels est plus stricte que celle de l'unidimensionnalité essentielle. En pratique, le modèle choisi s'ajuste souvent difficilement aux données car il est bien rare qu'en plus de la dimension dominante, une ou plusieurs dimensions mineures ne participent à l'organisation des réponses observées. On sait de plus que la stabilité des paramètres estimés est relativement sensible à la taille de l'échantillon et au nombre d'items. Les statistiques globales et analytiques accompagnant l'estimation des paramètres du MRI doivent donc toujours être interprétées avec prudence.

Nous prendrons pour exemple l'échelle de Transformation de Lettres dont les 14 items ont été construits en référence à un modèle cognitif dans lequel la difficulté de l'item, croissante du 1er au 14ème item, est définie *a priori* en fonction de la complexité de la règle et de la structure de la série (Butterfield, Nielsen, Tangen et Richardson, 1985). Conformément aux hypothèses de construction du matériel, on vérifie sur un échantillon aléatoire de 1000 sujets une baisse significative de la performance avec le rang de l'item dans l'échelle [$F_{exact}(13, 987)=27{,}24$; $p=0{,}000$]. La consistance interne de l'échelle mesurée par l'alpha de Cronbach est satisfaisante (0,89). L'amplitude moyenne des corrélations item-échelle (0,59) montre que les items ont une assez bonne puissance discriminative. On note enfin que les corrélations point-bisériales sont élevées et homogènes.

Tableau 2
Echelle de Transformation de Lettres : statistiques classiques (N = 1000).

Item	Probabilité de réussite *p*	Corrélation item/échelle	Corrélation point-bisériale
1	0,708	0,603	0,799
2	0,669	0,641	0,832
3	0,600	0,424	0,538
4	0,708	0,652	0,864
5	0,652	0,581	0,749
6	0,662	0,609	0,789
7	0,676	0,638	0,831
8	0,666	0,638	0,827
9	0,650	0,567	0,730
10	0,599	0,624	0,791
11	0,513	0,567	0,711
12	0,625	0,646	0,825
13	0,616	0,615	0,783
14	0,453	0,486	0,611

L'unidimensionnalité *hiérarchique* des données (Dickes, ce numéro) est testée en appliquant le modèle de Rasch pour lequel tous les items sont considérés également discriminants. On utilise le logiciel RSP (Glas et Ellis, 1993) qui propose plusieurs méthodes d'estimation de la difficulté b_j de l'item (CML: maximum de vraisemblance conditionnel ; MML: maximum de vraisemblance marginal) et de l'aptitude θ (ML : maximum de vraisemblance ; WML : maximum de vraisemblance pondéré ; EAP : méthode bayésienne avec distribution *a posteriori*). RSP fournit aussi plusieurs statistiques globales pouvant contribuer à évaluer l'adéquation du modèle aux données.

Une première statistique (R_0) est basée sur la différence entre la distribution des scores théoriques (pour les estimations par la méthode MML des paramètres de l'item et de la population) et celle des scores observés. Elle permet de tester l'hypothèse nulle de distribution normale du trait latent. La distribution asymptotique de R_0 étant connue, l'aptitude θ est distribuée normalement si cet indice est statistiquement non-significatif. Le postulat d'unidimensionnalité du modèle peut être testé à l'aide de statistiques de 1er ou de 2nd ordre calculées après avoir divisé l'échantillon de sujets en sous-groupes de score homogène, à partir des estimations fournies par la méthode MML (resp. CML) si le trait latent est (resp. n'est pas) distribué normalement. Les statistiques de 1er ordre (R_1 ou Q_1) permettent d'évaluer si les courbes caractéristiques de l'item (CCI) sont croissantes et ont la même forme logistique. Leur calcul est basé sur le dénombrement des réponses correctes à chacun des items dans chacun des sous-groupes. Leur valeur est fonction de l'amplitude des écarts entre le nombre observé et le nombre théorique d'individus de chacun des sous-groupes répondant correctement à l'item. Celle-ci est d'autant plus faible (au regard du nombre de degrés de liberté) que les CCI sont croissantes et de même forme logistique. Les statistiques de 2nd ordre (R_2 si le nombre d'items est inférieur à 15, Q_2 dans le cas contraire) sont sensibles au non-respect du postulat d'unidimensionnalité. Elles sont calculées à partir des écarts, pour chacun des sous-groupes de sujets, entre le nombre observé et le nombre théorique de personnes ayant simultanément bien répondu à deux items. Leur valeur doit donc être d'autant plus faible, pour un nombre donné de degrés de liberté, que l'hypothèse d'indépendance locale des items est bien respectée. Ces différents indices ayant une distribution asymptotique du χ^2, il est possible d'en tester la significativité.

Tableau 3
Items de Transformation de Lettres:
indices d'ajustement du modèle de Rasch (programme RSP) pour des échantillons aléatoires de 100, 200, 500 et 1000 sujets (utilisation de la méthode CML lorsque l'hypothèse de normalité du trait latent est réfutée).

Nombre de sujets	Méthode d'estimation	R_0	ddl	p	R_1	ddl	p	R_2	ddl	P
100	MML	12,294	12	0,4223	37,763	25	0,0488	159,556	102	0,0002
200	MML	38,463	12	0,0001						
	CML				57,402	39	0,0289	150,625	84	0,0000
500	MML	33,745	12	0.0007						
	CML				82,773	39	0,0001	138,314	84	0,0002
1000	MML	79,802	12	0.0000						
	CML				181,098	52	0,0000	361,398	84	0,0000

Les valeurs des statistiques obtenues en appliquant RSP à des échantillons de taille différente sont présentées dans le tableau 3. Les résultats des tests de significativité apparaissent clairement liés à la taille de l'échantillon. L'examen des probabilités associées à R_0 révèle en effet que l'hypothèse de normalité du trait latent n'est respectée que pour le seul échantillon de 100 sujets et que les probabilités associées à R_1 (postulat de monotonie et de similitude des CCI) ne sont non significatives que pour les échantillons de faible effectif. On est enfin conduit, quelle que soit la taille de l'échantillon, à rejeter l'hypothèse d'unidimensionnalité du trait latent. Malgré un nombre d'items un peu faible pour pouvoir considérer les estimations fournies comme vraiment satisfaisantes, l'adéquation du modèle de Rasch aux données recueillies avec TL n'est pas bonne. On peut néanmoins améliorer l'ajustement du modèle de Rasch en éliminant les items les plus " problématiques " que l'examen des statistiques locales (Q normalisés, U) permet de repérer. Bien que réduisant la fidélité de l'échelle, la suppression des items 3 et 14 conduit ainsi à une meilleure adéquation du modèle aux données [R_0=27,499 ; ddl=10 ; p=0,0022 ; R_1=43,314 ; ddl=33 ; p=0,108 ; R_2=70,863 ; ddl=60; p=0,159 ; méthode CML, N=500 sujets]. Mais il faut éliminer 6 items sur les 14 que comporte l'échelle pour que le modèle de Rasch s'ajuste à peu près correctement sur un échantillon de 1000 sujets. Il paraît donc raisonnable de conclure que l'hypothèse d'unidimensionnalité hiérarchique n'est qu'assez peu compatible avec l'organisation des données recueillies avec l'ensemble des items de l'échelle de Transformation de Lettres.

Qu'en est-il maintenant de la compatibilité de l'hypothèse d'unidimensionnalité *non hiérarchique* telle qu'elle est comprise dans les MRI unidimensionnels qui prennent en compte la puissance de discrimination des items (paramètre a_j de discrimination) et la possibilité de répondre correctement au hasard (asymptote basse c_j ou paramètre de pseudo-chance) ?

Pour tenter de répondre à cette question, nous avons utilisé le logiciel BILOG 3 (Mislevy et Bock, 1990) qui permet d'estimer les MRI unidimensionnels à plusieurs paramètres (calibration de l'item avec la méthode MML). Les modèles à 2 paramètres (a_j, b_j) et à 3 paramètres[12] (a_j, b_j, c_j) ont été appliqués aux réponses des 1000 sujets de l'échantillon précédent (échelle TL). Le nombre d'items étant ici inférieur à 20, BILOG 3 ne fournit pas de test statistique[13] permettant d'évaluer l'adéquation de ces modèles aux données. Les auteurs recommandent dans ce cas l'étude, item par item et en différents points de quadrature du trait latent θ, des " résidus postérieurs standardisés " (*Standardized Posterior Residuals*). Ces résidus standardisés sont calculés en chaque point de quadrature par différence entre la probabilité postérieure de réussite à l'item et la probabilité déduite du modèle correspondant. BILOG 3 fournit aussi une statistique moyenne d'ajustement de l'item (la racine carrée de la moyenne des carrés des résidus postérieurs : *RMR*). Globalement, une valeur de cette dernière supérieure à 2 est considérée comme témoignant d'un mauvais ajustement de l'item[14].

Les résultats obtenus en appliquant les modèles à 2 et 3 paramètres à l'ensemble des items de l'échelle sont présentés dans le tableau 4. On voit immédiatement que par rapport au modèle à 1 paramètre (valeur identique du paramètre de discrimination pour tous les items), c'est avec le modèle à 2 paramètres que le nombre d'items pour lesquels le *RMR* est supérieur à 2 est le plus faible. Les postulats posés sur les données par les modèles à 2 ou à 3 paramètres semblent donc mieux respectés que ceux du modèle de Rasch. La comparaison des estimations fournies pour les modèles à 2 et 3 paramètres est plus incertaine d'autant que le gain d'ajustement apporté par un modèle en comparaison à un autre ne peut être testé avec BILOG 3. On peut remarquer que le nombre d'items non conformes est un peu plus faible pour le modèle à 2 paramètres qu'il ne l'est pour celui à 3 paramètres et que la corrélation entre les paramètres de discrimination et de difficulté est plus importante pour le modèle à 2 paramètres ($r_{a_j b_j} = -0{,}484$)

[12] Quatre alternatives de réponse étaient en effet offertes aux sujets.

[13] BILOG 3 ne fournit un test du chi-deux du rapport de vraisemblance que lorsque l'analyse porte sur plus de 20 items.

[14] Cette valeur communément employée correspond approximativement à une erreur de type I de .05.

qu'elle ne l'est pour celui à 3 paramètres ($r_{a_j b_j} = -0{,}017$). Une évaluation plus fine de l'ajustement de ces deux modèles pourrait profiter de la comparaison de l'invariance des paramètres de l'individu (resp. de l'item) estimés sur des groupes d'items (resp. d'individus) différents ainsi que de l'examen, item par item et pour chaque modèle, des graphes des résidus standardisés.

Tableau 4
MRI unidimensionnels à 1, 2 et 3 paramètres : estimations fournies par BILOG 3 pour les 14 items de Transformation de Lettres (a_i : puissance discriminative ; b_i : difficulté ; c_i : asymptote basse ou paramètre de pseudo-chance ; *RMR* : résidu moyen) (N=1000).

Item	Modèle de Rasch (1 paramètre)		MRI à 2 paramètres			MRI à 3 paramètres			
	b_i	*RMR*	a_i	b_i	*RMR*	a_i	b_i	c_i	*RMR*
1	-0,816	1,453	1,153	-0,814	0,804	1,398	-0,527	0,146	0,918
2	-0,654	0,925	1,266	-0,637	1,416	1,392	-0,466	0,081	**2,322**
3	-0,388	**6,544**	*0,631*	-0,522	1,021	0,802	-0,060	0,179	1,453
4	-0,816	**4,483**	1,402	-0,769	1,115	1,539	-0,598	0,086	**2,453**
5	-0,587	0,893	1,032	-0,614	0,704	1,210	-0,363	0,118	1,388
6	-0,627	2,098	1,134	-0,633	1,890	1,236	-0,456	0,083	**2,378**
7	-0,682	**3,540**	1,231	-0,670	**3,348**	1,299	-0,527	0,067	**3,578**
8	-0,642	**2,727**	1,256	-0,627	1,569	1,344	-0,462	0,079	**2,447**
9	-0,580	**2,276**	0,967	-0,621	1,948	1,394	-0,182	0,202	0,685
10	-0,383	1,864	1,180	-0,388	1,117	1,491	-0,152	0,105	1,277
11	-0,068	**2,277**	1,013	-0,085	1,979	1,438	0,169	0,111	1,168
12	-0,483	**3,394**	1,295	-0,470	**2,282**	1,845	-0,169	0,140	0,553
13	-0,448	1,902	1,155	-0,454	1,584	1,546	-0,160	0,134	0,653
14	0,146	**3,840**	0,814	0,151	**2,634**	1,362	0,426	0,130	1,144

En gras, *RMR*>2 ; en italiques, a_i<0,8.

Les résultats apparaissant dans le tableau 4 montrent que l'hypothèse d'unidimensionnalité non hiérarchique est plus compatible avec les réponses observées avec l'échelle de Transformation de Lettres dont seraient par exemple éliminés les items 7, 12 et 14 que ne l'est celle d'unidimensionnalité hiérarchique. On voit néanmoins que l'application d'un MRI unidimensionnel doit souvent s'accompagner de l'élimination, souvent problématique (Flieller, 1994), d'un nombre plus ou moins important d'items pour pouvoir espérer parvenir au respect des postulats posés par le modèle et aboutir, par une sorte de mise à l'épreuve continuée, à une représentation certes unidimensionnelle mais s'appliquant à un nom-

bre réduit d'items sélectionnés empiriquement. L'intérêt de ces modèles unidimensionnels réside donc moins dans la possibilité offerte de juger la dimensionnalité d'un ensemble donné d'items que dans celle d'atteindre empiriquement l'unidimensionnalité par élimination d'items non conformes ou par regroupement d'items adéquats en sous-groupes plus homogènes.

Utilisation des modèles multidimensionnels

L'étude de la dimensionnalité de données dichotomiques peut aussi et surtout bénéficier de l'utilisation de l'analyse factorielle d'items et des MRI multidimensionnels. L'intérêt de ces approches multidimensionnelles sera illustré par un exemple qui utilise les réponses des sujets à 24 items du test de raisonnement (8 items de LO, 8 items de RS et 8 items de TL) choisis pour avoir conduit à un taux de réussite intermédiaire. Deux échantillons aléatoires de 1000 sujets sont constitués. L'hypothèse que l'on souhaite vérifier et que le graphique des valeurs propres de la matrice des corrélations tétrachoriques estimées à partir des données de l'échantillon 1 (figure 2) ne semble d'ailleurs pas contredire, est celle de l'existence d'une dimension par échelle.

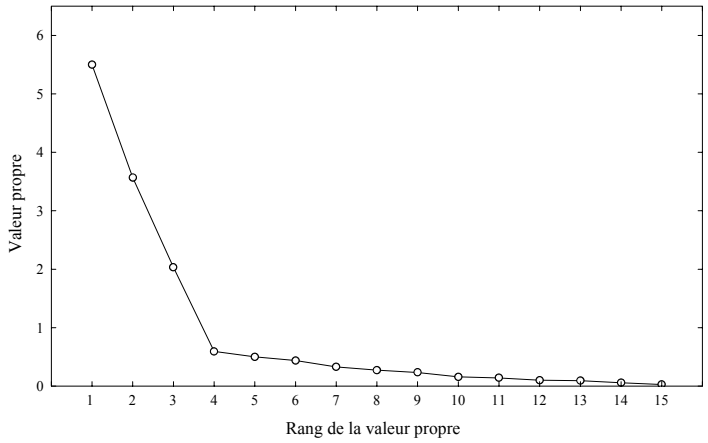

Figure 2
Graphique des valeurs propres de la matrice de corrélations tétrachoriques entre 24 items du test de raisonnement (échantillon 1, N=1000).

On applique d'abord aux données de l'échantillon 1 le modèle d'analyse factorielle non linéaire de McDonald (programme NOHARM, Fraser et McDonald, 1988). NOHARM estimant les paramètres pour le modèle ogive normale à 2 paramètres, les valeurs de c_i sont d'abord calculées au moyen de BILOG 3 puis

fixées dans le modèle. Plusieurs analyses confirmatoires sont ensuite réalisées sur les données du 1er échantillon afin d'estimer les paramètres de la solution unidimensionnelle, de la solution à 3 facteurs orthogonaux LO, RS et TL et d'une solution à 3 facteurs (LO, LO-RS-TL, TL) suggérée par les résultats d'analyses effectuées avec DIMTEST montrant que LO est dimensionnellement distincte de RS et TL (T=8,763 ; p=0,000), que TL est dimensionnellement distincte de LO et RS (T=6,909 ; p=0,000) mais que RS n'est pas dimensionnellement distincte de LO et TL (T=0,585 ; p=0,279).

Tableau 6

Analyse factorielle non linéaire confirmatoire d'items (NOHARM) (LO : items de Logique Opératoire ; RS : items de Raisonnement Spatial ; TL : items de Transformation de Lettres ; RMR : moyenne des covariances résiduelles ; échantillon 1 ; N = 1000).

Nombre de dimensions	Items correspondants	Nombre d'itérations	RMR
1	LO-RS-TL	111	0,02356
3	LO, RS, TL	80	0,01981
3	LO, RS-TL, TL	9	**0,01866**

Les informations présentées dans le tableau 6 résument au plan de l'ajustement global les résultats de ces analyses. La meilleure solution est identifiée à partir de l'examen de la matrice des covariances résiduelles inter-items. L'adéquation du modèle est ici d'autant meilleure que le RMR (*Root Mean Square Residual*), une mesure de la moyenne des covariances résiduelles, est faible. On voit clairement que la solution unidimensionnelle présente un moins bon ajustement que les solutions multidimensionnelles parmi lesquelles la solution faisant l'hypothèse de trois dimensions respectivement mesurées par les items de LO, par ceux de RS et Tl et par ceux de TL, s'avère être la plus adéquate. La comparaison paire d'items par paire d'items entre covariances résiduelles et covariances observées met en outre en lumière un problème d'ajustement spécifique aux items de LO. La moyenne des covariances résiduelles entre ces seuls items est en effet très élevée au regard de celle des covariances observées (0,049 *vs* 0,066) alors que ce résultat n'est pas observé lorsqu'on considère par exemple les seuls items de TL (0,0001 *vs* 0.085). Ce défaut d'adéquation (le RMR est largement supérieur à l'erreur-type des résidus[15]) montre donc que contrairement aux items de TL, les items de LO ne satisfont pas à l'exigence d'indépendance locale[16] et ne pourraient

[15] Ici $1/\sqrt{N} = 0{,}0268$.

[16] Ce résultat se comprend bien quand on sait que les items de LO ont été construits en référence à un modèle cognitif dans lequel la difficulté de l'item croît progressivement en fonction de règles que l'individu apprend à maîtriser au cours de la tâche.

donc être ajustés de manière satisfaisante par un MRI unidimensionnel non dépendant localement.

Tableau 7

Modèle à 3 dimensions orthogonales: estimations fournies par les programmes NOHARM (paramétrisation en facteurs communs : seuil γ_j, saturations λ_{j1}, λ_{j2}, et λ_{j3}, unicité ψ_j ; et TESTMAP (paramétrisation en traits latents : paramètres a_{j1}, a_{j2} et a_{j3} de discrimination ; paramètre d_j de difficulté de l'item)
(p : probabilité observée de réussite à l'item ; échantillon 2 ; N=1000).

		Modèle NOHARM Paramétrisation en facteurs communs					MRI multidimensionnel à 2 paramètres					
Item	p	γ_j	λ_{j1}	λ_{j2}	λ_{j3}	ψ_j	d_j	a_{j1}	a_{j2}	a_{j3}	$MDISC_j$	$MDIFF_j$
1_{LO}	0,692	-0,295	0,984	-	-	0,032	0,764	0,938	-	-	0,938	-0,814
2_{LO}	0,462	-1,458	0,984	-	-	0,031	-0,077	0,617	-	-	0,617	0,125
3_{LO}	0,600	-0,842	0,985	-	-	0,029	0,422	0,940	-	-	0,940	-0,449
4_{LO}	0,572	-1,033	0,986	-	-	0,028	0,349	1,044	-	-	1,044	-0,334
5_{LO}	0,398	-1,399	0,986	-	-	0,029	-0,292	0,802	-	-	0,802	0,364
6_{LO}	0,419	-1,327	0,986	-	-	0,027	-0,230	0,887	-	-	0,887	0,259
7_{LO}	0,431	-1,325	0,986	-	-	0,027	-0,187	0,878	-	-	0,878	0,213
8_{LO}	0,355	-1,618	0,984	-	-	0,037	-0,382	0,497	-	-	0,497	0,769
9_{RS}	0,852	0,833	-	0,628	-	0,605	1,127	-	0,394	-	0,394	-2,860
10_{RS}	0,521	-0,922	-	0,896	-	0,197	0,056	-	0,238	-	0,238	-0,235
11_{RS}	0,817	0,680	-	0,782	-	0,389	0,971	-	0,400	-	0,400	-2,428
12_{RS}	0,452	-0,986	-	0,979	-	0,042	-0,121	-	0,524	-	0,524	0,231
13_{RS}	0,518	-0,999	-	0,982	-	0,037	0,070	-	0,629	-	0,629	-0,111
14_{RS}	0,382	-1,084	-	0,981	-	0,037	-0,341	-	0,657	-	0,657	0,519
15_{RS}	0,386	-0,971	-	0,985	-	0,029	-0,429	-	1,124	-	1,124	0,382
16_{RS}	0,378	-1,297	-	0,984	-	0,033	-0,371	-	0,750	-	0,750	0,495
17_{TL}	0,727	0,460	-	0,313	0,823	0,226	0,904	-	0,491	1,036	1,146	-0,789
18_{TL}	0,691	0,400	-	0,384	0,790	0,228	0,763	-	0,538	1,088	1,214	-0,629
19_{TL}	0,745	0,524	-	0,402	0,870	0,082	1,106	-	0,528	1,251	1,358	-0,815
20_{TL}	0,697	0,418	-	0,424	0,766	0,234	0,797	-	0,548	1,098	1,227	-0,649
21_{TL}	0,710	0,341	-	0,366	0,794	0,236	0,798	-	0,478	0,973	1,084	-0,736
22_{TL}	0,672	0,377	-	0,092	0,958	0,075	0,725	-	0,500	1,259	1,355	-0,535
23_{TL}	0,677	0,352	-	0,287	0,768	0,328	0,644	-	0,352	1,008	1,068	-0,603
24_{TL}	0,503	-0,098	-	0,304	0,655	0,479	-0,031	-	0,313	0,794	0,853	0,036

Nous avons enfin, comme le suggèrent par exemple McDonald (1997) et Reckase (1997a), tenté de valider la structure dimensionnelle identifiée en répliquant ces analyses sur un échantillon équivalent de sujets. Des résultats tout à fait semblables étant observés sur ce second échantillon, nous avons retenu la solution à 3 facteurs orthogonaux (LO, RS-TL, TL) dont les paramètres apparaissent dans

le tableau 7. La paramétrisation en facteurs communs, dérivée de celle en traits latents non présentée ici, permet d'interpréter ces paramètres aisément (McDonald, 1997). Le paramètre de seuil γ_j est la transformation normale inverse du paramètre classique *p* de difficulté de l'item : la valeur du seuil est donc d'autant moins élevée que l'item est difficile (la corrélation entre γ_j et *p* est de 0,922). Quant aux saturations λ_{jk} et aux unicités, elles s'interprètent comme habituellement en analyse factorielle. On remarque ainsi que la 2nde dimension est bien mesurée par les items de RS et dans une moindre mesure par ceux de TL (à l'exception de l'item 22) ou que les items de TL sont dimensionnellement plus complexes que ceux de LO qui saturent uniquement et extrêmement fortement la 1ère dimension. L'examen des unicités permet par ailleurs de repérer les items dont la variance non prise en compte par les dimensions du modèle est importante (e.g., les items 9, 10, 11 et la très grande majorité des items de TL).

Cette même démarche confirmatoire a guidé l'application des MRI multidimensionnels à 2 et 3 paramètres (programme TESTMAP, McKinley, 1992) dont nous avons dit qu'ils apportent des informations spécifiques du point de vue des caractéristiques des items. Plusieurs modèles faisant les mêmes hypothèses dimensionnelles que ceux testés précédemment avec NOHARM ont donc été appliqués aux données de l'échantillon 2. L'adéquation de chacun de ces modèles est évaluée avec le CAIC, une statistique dérivée du logarithme de la vraisemblance de la solution et interprétable en termes de proximité du modèle par rapport au modèle vrai. L'adéquation du modèle est donc d'autant meilleure que la valeur du CAIC est plus faible. Les résultats obtenus sont conformes aux constatations effectuées avec NOHARM ; la solution qui présente la meilleure adéquation et pour laquelle aucun problème de convergence n'est rencontré est celle à 3 dimensions LO, RS-TL et TL, le MRI à 2 paramètres s'ajustant globalement un peu mieux que celui à 3 paramètres (tableau 8).

Tableau 8

Valeurs du CAIC (*Consistent Akaike Information Criterion*) pour chacun des MRI appliqués aux données de l'échantillon 2 (N = 1000).

Nombre de dimensions	Items correspondants	MRI multidimensionnel à 2 paramètres	MRI multidimensionnel à 3 paramètres
1	LO-RS-TL	28612,29	28636,48
3	LO, RS,TL	27107,05	27210,56
3	LO, RS-TL,TL	**27029,68**	27132,04

Nous avons présenté les estimations fournies par TESTMAP et certaines statistiques dérivées dans le même tableau que celles fournies par NOHARM (tableau 7) afin de faciliter la comparaison des résultats obtenus avec les deux

méthodes. Les corrélations entre les *k* dimensions θ étant nulles, les paramètres a_{jk} rendent compte à eux seuls des corrélations entre les scores aux items. Ces paramètres de discrimination peuvent être interprétés pour chaque dimension comme ils le sont dans les MRI unidimensionnels (Reckase, 1997a, Ackerman, 1996). Ils indiquent la sensibilité de l'item aux différences d'aptitude sur le continuum latent correspondant. On remarque ainsi qu'à l'exception des items 15 et 16, les items de l'échelle RS ont une faible puissance discriminative ou que les items 8 et 2 de l'échelle LO sont insuffisamment discriminants. Un indicateur plus global ($\mathrm{MDISC}_j = \sqrt{\sum a_{jk}^2}$) permet de mesurer, pour la meilleure combinaison d'aptitudes, la puissance discriminative des items multidimensionnels. C'est le cas des items de l'échelle TL, globalement d'une puissance discriminative élevée mais logiquement plus sensibles aux différences d'aptitude par rapport à la 3ème dimension qu'ils ne le sont par rapport à la 2nde.

Par ailleurs et ainsi que le souligne (Reckase, 1997a), l'interprétation du paramètre d_j lié à la difficulté de l'item ne peut se faire comme celle du paramètre b_j dans les MRI unidimensionnels. Une statistique interprétable comme b_j ($\mathrm{MDIFF_j} = \frac{-d_j}{\mathrm{MDISC_j}}$) peut néanmoins être calculée ; celle-ci indique la distance entre l'origine de l'espace latent et le point où la pente de la surface de réponse à l'item est la plus prononcée. A une valeur élevée de cette statistique correspondent donc un niveau élevé de difficulté de l'item et une faible probabilité de réponse à l'item (la distribution de $\mathrm{MDIFF_j}$ présente une corrélation de -0,91 avec celle de *p*).

Au total, l'utilisation de ces deux approches multidimensionnelles dans l'étude de la dimensionnalité des données dichotomiques recueillies avec les trois échelles du test de raisonnement aboutit à l'identification d'une même structure dimensionnelle. Du point de vue de l'analyse des caractéristiques des items, la comparaison des résultats obtenus montre l'existence d'une forte relation entre les valeurs du paramètre de seuil γ_j de l'analyse factorielle non linéaire et celles de l'indicateur global de difficulté du MRI multidimensionnel à 2 paramètres (corrélation de -0,800 entre γ_j et $\mathrm{MDIFF_j}$). L'observation d'une corrélation non négligeable (0,374) entre γ_j et $\mathrm{MDISC_j}$ et de corrélations moyennes entre les saturations λ_{jk} (souvent très élevées) et les paramètres de discrimination a_{jk} (resp. 0,581, 0,442 et 0,900 pour les dimensions 1, 2 et 3) amène cependant à préférer la se-

conde approche qui semble mieux distinguer les paramètres de discrimination et de difficulté ($r(\text{MDIFF}_j, \text{MDISC}_j) = 0{,}057$) des 24 items considérés et faciliter par là-même l'analyse de leurs propriétés psychométriques. Une fois identifiée la structure dimensionnelle des données et effectué le calibrage multidimensionnel des items, les scores individuels peuvent être estimés pour chacune des trois dimensions retenues (e.g., au moyen du programme THSCORE de Ferrando et Lorenzo, 1998).

CONCLUSION

Beaucoup de psychologues, assez réticents à l'égard des postulats qui fondent l'emploi des MRI unidimensionnels (e.g., Reuchlin, 1997), limitent souvent l'usage de ces modèles à la construction d'ensembles d'items homogènes, par exemple les items d'un test de vocabulaire, dont le caractère d'unidimensionnalité est vérifié grâce aux MRI unidimensionnels. Les développements auxquels l'étude des MRI a conduit ces vingt dernières années offrent pourtant des perspectives, tant au plan théorique qu'au plan des applications, qu'il serait dommage de négliger (e.g., De Boeck et Van Mechelen, ce numéro). Nous avons pour notre part essayé d'illustrer l'intérêt de certaines extensions multidimensionnelles de ces modèles dans l'abord d'un problème important en psychologie différentielle, celui de la structure dimensionnelle d'un ensemble d'items, ici dichotomiques, administrés à une catégorie donnée d'individus.

Nous avons utilisé dans ce travail plusieurs méthodes basées sur les MRI paramétriques et non paramétriques. Les conclusions auxquelles peuvent conduire les résultats obtenus sont bien évidemment fonction du type d'objectif fixé et du niveau d'analyse privilégié. Elles dépendent aussi de la définition donnée de l'unidimensionnalité, des caractéristiques spécifiées du modèle sous-jacent et des techniques d'estimation employées. Aucune procédure n'étant unanimement acceptée pour évaluer la structure dimensionnelle de données dichotomiques, l'utilisation de plusieurs méthodes d'analyse s'impose donc le plus souvent, l'utilisation dans une démarche confirmatoire de confrontation entre une modélisation spécifiée *a priori* et les données d'observation étant la plus heuristique. Dans cette perspective et malgré les difficultés associées à leur emploi, les modèles multidimensionnels sur lesquels nous avons plus particulièrement insisté dans ce travail nous paraissent offrir d'utiles perspectives, tant du point de vue de l'identification de la structure dimensionnelle des données que de celui de l'analyse psychométrique des items permettant la mesure multidimensionnelle des différences individuelles.

BIBLIOGRAPHIE

Ackerman, T. (1996). Graphical representation of multidimensional IRT analyses. *Applied Psychological Measurement, 20*, 311-329.

Bernstein, I.H., & Teng, G. (1989). Factoring items and factoring scales are different : Spurious evidence for multidimensionality due to item categorisation. *Psychological Bulletin, 105*, 467-477.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters : An application of the EM algorithm. *Psychometrika, 46*, 433-449.

Bock, R.D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Butterfield, E., Nielsen, D., Tangen, K.L., & Richardson, M.B. (1985). Theoretically based psychometric measures of inductive reasoning. In S.E. Embretson (Ed.), *Test design : Developments in psychology and psychometrics* (pp. 77-147). Orlando, Academic Press.

Carroll, J.B. (1945). The effects of difficulty and chance success on correlations between items and between tests. *Psychometrika, 26,* 347-372.

De Champlain, A., & Linda Tang, K. (1997). CHIDIM* : A Fortran program for assessing the dimensionality of binary item responses based on McDonald's nonlinear factor analytic model. *Educational and Psychological Measurement, 57*, 174-178.

Dickes, P. (1996). L'analyse factorielle et ses deux logiques d'application. *Psychologie Française, 41*, 9-22.

Dickes, P., Flieller, A., Tournois, J. & Kop, J.-L. (1994). *La psychométrie.* Paris : PUF.

Ferrando, P.J. & Lorenzo, U. (1998). TH-SCORE*: A program for obtaining ability estimates under different psychometric models. *Educational and Psychological Measurement, 58*, 841-845.

Flieller, A. (1994). Méthodes d'étude de l'adéquation au modèle logistique à un paramètre (modèle de Rasch). *Mathématiques, Informatique et Sciences Humaines, 127*, 19-47.

Fraser, C., & McDonald, R.P. (1988). NOHARM* : Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267-269.

Glas, C.A.W., & Ellis, J.J. (1993). *Rasch Scaling Program (RSP): User's manual.* ProGamma, Groningen : The Netherlands.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York : Macmillan.

Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory : Principles and applications.* Boston : Kluwer.

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

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. London : Sage Publications.

Hattie, J. (1985). Methodological review : Assessing unidimensionality for tests and items. *Applied Psychological Measurement, 9*, 139-164.

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

Hoskens, M. & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods, 2*, 261-275.

Laveault, D., Zumbo, B., Gessaroli, M.E., & Boss, M.W. (Eds.). *Modern theories of measurement : Problems and issues*. Ottawa : Edumetrics Research Group.

Lord, F. (1980). A*pplications of Item Response Theory to practical testing problems*. Hillsdale, N.J. : LEA.

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

McDonald, R.P. (1967). Non linear factor analysis. *Psychometric Monographs,* N° 15, 1-167.

McDonald, R.P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*, 379-396.

McDonald, R.P. (1997). Normal-Ogive Multidimensional Model. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item repose theory* (pp. 257-269). New York : Springer Verlag.

McKinley, R.L. (1988). *Assessing dimensionality using confirmatory multidimensional IRT*. Paper presented at the Annual Meeting of the American Educational Research.

McKinley, R.L. (1989). *Confirmatory analysis of test structure using multidimensional IRT*. Research report 89-31, Educational Testing Service, Princeton, NJ.

McKinley, R.L. (1992). *TestMap 2.1* : User's guide*. Princeton, NJ : ETS.

McKinley, R.L., & Reckase, M. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space*. Research report ONR83-2. Iowa city, IA : ACT.

Mislevy, R.J., & Bock, R.D. (1990). *BILOG 3 : Item analysis and test scoring with binary logistic models*. Chicago: Scientific Software.

Mokken, R.J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417-430.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 99-117.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement, 17*, 29-38.

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

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen Danish Institute For Educational Research). Chicago, The University of Chicago Press.

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

Reckase, M. (1997a). A linear logistic multidimensional model for dichotomous item response model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York : Springer.

Reckase, M. (1997b). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25-36.

Reuchlin, M. (1997). *La psychologie différentielle.* Paris: PUF.

Sijtsma, K. (1998). Methodology review : Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22*, 3-31.

Stout, W. (1987). A statistical approach for determining the latent trait dimensionality in psychological testing. *Psychometrika, 55*, 293-326.

Stout, W., Douglas, J., Junker, B., & Roussos, L. (1993). *DIMTEST* Manual.* Department of Statistics, University of Illinois at Urbana-Champaign.

van der Linden, W. & Hambleton, R. (Eds.), *Handbook of modern item response theory.* New York : Springer Verlag.

Verhelst, N.D., & Glas, C.A.W. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58*, 395-415.

Wilson, D., Wood, R., & Gibbons, R.D. (1991). *TESTFACT : Test scoring, item statistics, and item factor analysis.* Chicago : Scientific Software.

Wilson, M. (1992) (Ed.). *Objective measurement : Theory into practice* (Vol. 1). Greenwich : Ablex.

Wilson, M. (1994) (Ed.). *Objective measurement : Theory into practice* (Vol. 2). Greenwich : Ablex.

Wilson, M., Engelhard, G. & Draney, K. (Eds.) (1997). *Objective measurement : Theory into practice* (Vol. 4). Greenwich : Ablex.

Wright, B.D., & Stone, M.H. (1979). *Best test design.* Chicago : MESA Press.

Note: Les programmes marqués d'une astérisque peuvent être obtenus gratuitement auprès des auteurs.