INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Université d'Ottawa · University of Ottawa

Étude sur la fidélité et l'efficacité relatives de méthodes d'analyse du fonctionnement différentiel des items applicables à des échantillons de taille réduite

Thèse présentée à
l'École des études supérieures et de la recherche
de l'Université d'Ottawa
afin de satisfaire à l'une des exigences
du programme de doctorat en éducation

Université d'Ottawa

1998

REMERCIEMENTS

Cette thèse n'aurait pu voir le jour sans l'aide et le soutien de plusieurs personnes, et je les en remercie.

Je voudrais dire un merci tout particulier au Dr. Dany Laveault, mon directeur de thèse, dont la patience, les encouragements et les conseils m'ont permis de mener ce projet à terme. Je voudrais aussi dire merci au Dr. Bruno Zumbo, au Dr. Marc Gessaroli et au Dr Marvin Boss, pour leurs conseils, au moment de choisir un sujet. Je voudrais enfin remercier le Dr. Michel Brabant pour son aide lors de la cueillette et du traitement des données.

Je voudrais aussi remercier mes parents pour avoir fait leur ce projet et avoir cherché à me faciliter la tâche de mille et une façons. Je voudrais surtout les remercier de m'avoir enseigné le travail et la persévérance. Sans les valeurs qu'ils m'ont inculquées, je n'aurai pu mener ce projet à terme.

RÉSUMÉ

La présente étude examine la fidélité et l'efficacité de la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995) pour étudier le fonctionnement différentiel des items avec des échantillons de taille réduite. Elle examine aussi la fidélité et l'efficacité de la méthode du khi carré et du delta de Mantel-Haenszel et de la méthode de régression logistique. Ces dernières ont été retenues parce qu'elles se classent parmi les plus prometteuses des méthodes applicables à des échantillons de taille réduite. La fidélité des méthodes est évaluée par la stabilité de leurs indices ; leur efficacité, par leur capacité de déceler des items qui ont un fonctionnement différentiel d'une grandeur prédéterminée. À titre complémentaire, nous avons également examiné la validité et la fidélité des décisions prises à partir des indices fournis par chaque méthode. Pour ce faire, nous avons considéré le nombre de fois que chaque item est détecté, la relation entre les items détectés dans les échantillons et les items reconnus comme ayant un fonctionnement différentiel et le nombre de décisions correctes. La validité et la fidélité des décisions et l'efficacité des méthodes ont été examinées pour divers critères de décision.

L'examen de la fidélité et de l'efficacité des méthodes repose sur des données réelles, les résultats à un test de mathématiques de l'American College Testing Program (ACTP). Pour étudier l'efficacité des méthodes, nous avons déterminé quels items avaient un fonctionnement différentiel non négligeable. Pour ce faire, nous nous sommes servis d'un échantillon de 40 000 sujets qui constitue, à toutes fins utiles, la population d'origine. Les hommes y forment le groupe de référence ; les femmes, le groupe focalisé. Ont été identifiés comme ayant un

fonctionnement différentiel non négligeable tous les items dont la différence de difficulté standardisée est égale ou supérieure à 0,05 en valeur absolue, après arrondissement au centième le plus proche. Les items devaient également présenter un delta de Mantel-Haenszel ou une différence de difficulté standardisée modifiée qui les rangent parmi les items avec les indices les plus grands. De plus, ils devaient présenter des différences de difficulté standardisée qui sont plus grandes que celles observées à l'intérieur du groupe de référence et du groupe focalisé.

Pour étudier la fidélité et l'efficacité des méthodes, nous avons sélectionné des échantillons de 250, de 500, de 1000 et de 2000 sujets parmi les 40 000 sujets représentant la population. Dans chaque cas, nous avons prélevé un nombre égal d'hommes et de femmes, de sorte que les échantillons se partagent en deux groupes égaux. La sélection des sujets a été effectuée de façon aléatoire et indépendante. L'échantillonnage a été répété 100 fois pour chaque grandeur d'échantillon.

À partir des analyses effectuées, il appert que la méthode de standardisation modifiée offre des indices dont la stabilité est un peu moins grande que celle des deltas de Mantel-Haenszel. Malgré cela, sa capacité de détecter des items qui ont un fonctionnement différentiel se compare avantageusement à la méthode du delta de Mantel-Haenszel. Quelle que soit la méthode à l'étude, la stabilité des indices et leur capacité de détecter des items qui ont un fonctionnement différentiel non négligeable varient en fonction de la taille des échantillons. Règle générale, stabilité et efficacité tendent à augmenter avec l'accroissement du nombre de sujets dans les échantillons.

Toutes les méthodes se comportent de façon analogue lorsqu'on se base sur des centiles pour déterminer quels items ont un fonctionnement différentiel dans les échantillons de taille réduite. Il en résulte peu de différence entre les méthodes pour des échantillons de même grandeur. Lorsqu'on se base sur des critères fixes, la méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique affichent des tendances à l'opposé de la méthode du delta de Mantel-Haenszel et de la méthode de standardisation modifiée. Cette divergence résulte de la nature des indices fournis et se traduit par des différences importantes dans la détection des items. De fait, l'inefficacité de la méthode du khi carré de Mantel-Haenszel et de la méthode de régression logistique tient à leur propension à identifier un nombre croissant d'items avec des différences minimes lorsque la grandeur des échantillons augmente. L'inefficacité de la méthode du delta de Mantel-Haenszel et de la méthode de standardisation modifiée tient à l'imprécision de leurs indices, imprécision qui se traduit par une propension à identifier comme ayant un fonctionnement différentiel un nombre d'autant plus grand d'items que les échantillons sont plus petits.

À partir des données recueillies, il appert que des échantillons de 250 sujets sont insuffisants pour assurer la stabilité des indices. Des échantillons de 500 sujets sont problématiques, des échantillons de 1000 sujets semblent un minimum et des échantillons de 2000 sujets sont préférables. Par contre, il est possible de détecter un nombre minimalement acceptable d'items qui ont un fonctionnement différentiel d'une grandeur prédéterminée avec des échantillons de 500 à 1000 sujets.

TABLE DES MATIÈRES

LISTE DE TABLEAUX

LISTE DES FIGURES

CHAPITRE I

INTRODUCTION

L'étude du fonctionnement différentiel des items (FDI) a pour but de déterminer si des groupes distincts de personnes répondent de la même façon aux différents items qui composent un test ou si des différences existent (Angoff, 1988). La plupart des recherches sur le fonctionnement différentiel des items se concentrent sur l'identification des items potentiellement biaisés. Employée dans ce but, l'étude du fonctionnement différentiel des items assure que les scores ont la même signification pour tous les sujets et que des facteurs non pertinents ne donnent pas un avantage indu à certains groupes de personnes au détriment des autres. En principe, un groupe d'items n'a pas de biais si les items sont influencés par les mêmes sources de variation dans tous les groupes de la population et que la distribution des sources de variation non pertinentes est la même chez des personnes d'un même niveau d'habileté au regard du concept mesuré (Crocker et Algina, 1986, p. 377). Il y a biais si l'une ou l'autre de ces conditions n'est pas satisfaite.

Traduite en termes statistiques, la notion de biais suppose une relation entre les réponses aux items et l'appartenance à un groupe chez des personnes d'un même niveau d'habileté. Un item n'est pas biaisé si, à un même niveau d'habileté, les réponses dépendent uniquement de l'habileté des sujets et non de l'effet combiné de leur habileté et de leur appartenance à un groupe. Il y a biais s'il existe une dépendance conditionnelle entre les réponses à l'item et l'appartenance à un groupe (Mellenbergh, 1989, p. 129 ; Mellenbergh et Kok, 1991, p. 292). Le biais est expliqué si l'introduction d'une variable a pour effet de réduire ou d'éliminer cette dépendance (Mellenbergh et Kok, 1991, p. 299-301).

Appliquée à des items notés de façon dichotomique, l'existence de biais dans un item se traduit par une différence dans la probabilité de répondre correctement à l'item chez des sujets d'un même niveau d'habileté qui appartiennent à des groupes distincts de la population. On y distingue deux types de fonctionnement différentiel (FD) : un fonctionnement différentiel uniforme (FDU) et un fonctionnement différentiel non uniforme (FDNU). Le premier correspond à une différence dans la probabilité de répondre correctement à un item qui est constante à tous les niveaux d'habileté au regard du concept mesuré ; le second, à une différence qui varie en fonction du niveau d'habileté des sujets (Mellenbergh, 1982, p. 115). La figure 1, à l'annexe A, illustre les deux types de fonctionnement différentiel.

Au cours des années, on a proposé plusieurs méthodes statistiques pour étudier le fonctionnement différentiel des items de tests composés de questions à choix multiple et ainsi détecter les items potentiellement biaisés. Les méthodes les plus souvent utilisées postulent l'unidimensionnalité des items (Shepard, 1982 ; Osterlind, 1983 ; Scheuneman et Bleistein, 1989 ; Mellenbergh et Kok, 1991 ; Ackerman, 1992). Certaines, plus récentes, prennent en compte la multidimensionnalité des items (Ackerman, 1992), mais elles exigent des échantillons de plusieurs milliers de personnes et des programmes d'ordinateur nettement plus complexes, ce qui en restreint l'utilisation.

Parmi les méthodes qui postulent l'unidimensionnalité des items, les méthodes basées sur la théorie des réponses aux items (TRI) et le modèle à trois paramètres sont considérées comme les plus appropriées au plan théorique (Shepard, Camilli et Averill, 1981 ; Mellenbergh, 1982 ; Shepard, Camilli et Williams, 1985 ; Kok, Mellenbergh et Van der Flier, 1985 ; McCauley et Mendoza 1985 ; Scheuneman et Bleistein, 1989 ; Park et Lautenschlager, 1990 ; Angoff, 1993).

Outre le fait que le procédé est complexe et onéreux et que son efficacité dépend du type d'indices retenu, des échantillons d'au moins 1000 personnes par groupe sont nécessaires pour obtenir une estimation précise des paramètres d'item. Réunir un aussi grand nombre de sujets pour s'assurer de la comparabilité des scores pour tous les groupes de la population à qui un test est destiné n'est guère pratique et pas toujours réalisable.

Plusieurs méthodes ne requièrent pas l'estimation de paramètres. Mis à part la méthode de standardisation qui nécessite des échantillons d'au moins 5000 sujets par groupe (Hills, 1989, p. 9), ces méthodes peuvent s'appliquer à des échantillons de moins de 1000 sujets par groupe. Parmi ces dernières, la méthode de Mantel-Haenszel et la méthode logit sont sans conteste les plus prometteuses. Bien qu'elle nécessite l'estimation de paramètres, la méthode de régression logistique peut également s'appliquer à des échantillons de moins de 1000 sujets par groupe. Les recherches qui portent sur ces méthodes tendent cependant à démontrer une relation entre la grandeur des groupes comparés et le taux de détection des items potentiellement biaisés, une perte d'efficacité d'autant plus grande et une généralisabilité des résultats d'autant plus mal assurée que les échantillons comptent peu de sujets.

Tout récemment, Ramsay (1989, 1991 a 1991 b, 1992 a, 1992 b, 1993 et 1995) a proposé un modèle non paramétrique de réponses aux items pour estimer la probabilité des réponses en fonction de l'habileté des sujets. La méthode procède par lissage et ne nécessite pas autant de sujets que les modèles paramétriques. Selon Dorans, Schmitt et Bleistein (1992, p. 311), l'utilisation des probabilités ainsi estimées et d'un procédé analogue à la méthode de standardisation pourrait constituer une solution de rechange acceptable pour étudier le fonctionnement différentiel des items avec des échantillons de taille réduite. Ramsay lui-même

étudie la possibilité d'utiliser l'estimation des probabilités obtenues à l'aide de son modèle avec la méthode de standardisation (Dorans, Potenza et Ramsay, 1993).

L'efficacité du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995) combiné à la méthode de standardisation et la généralisabilité des résultats par rapport à d'autres méthodes d'analyse du fonctionnement différentiel des items applicables à des échantillons de moins de 1000 sujets par groupe n'ont fait l'objet d'aucune recherche. La présente étude a pour but d'examiner la question et de vérifier l'effet dû à la grandeur des groupes. Dans les pages qui suivent, nous passerons en revue les principales méthodes d'analyse du fonctionnement différentiel des items applicables à des échantillons de taille réduite et nous verrons comment le modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995) peut s'appliquer à l'étude du fonctionnement différentiel des items. Nous ferons ensuite le point sur l'état de la recherche en ce qui concerne l'efficacité des autres méthodes et la généralisabilité des indices d'un échantillon à un autre.

CHAPITRE II

REVUE DE LA LITTÉRATURE

Description de méthodes d'analyse du fonctionnement différentiel des items (FDI)

Outre le modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993, 1995) combiné à la méthode de standardisation, les méthodes d'analyse du fonctionnement différentiel des items susceptibles de s'appliquer à des échantillons de moins de 1000 sujets par groupe comprennent les méthodes basées sur l'analyse de la variance, sur les indices de difficulté ou sur les indices de discrimination, les méthodes basées sur les modèles à un ou à deux paramètres, les méthodes basées sur des tableaux de contingence et les méthodes basées sur la régression. Toutes ces méthodes postulent une échelle commune sur laquelle les scores de chaque groupe peuvent être comparés et elles vérifient les corollaires de ce postulat (Van de Vijver et Poortinga, 1991, p. 286).

Les méthodes proposées se partagent en deux grandes catégories : les méthodes conditionnelles et les méthodes non conditionnelles (Mellenbergh, 1982, p. 106 ; 1989, p. 128 ; Van der Flier, Mellenbergh, Ader et Wijn, 1984, p. 131-132). Une méthode est dite conditionnelle si elle vérifie l'existence de différences dans le fonctionnement d'un item à divers niveaux d'habileté, de manière à prendre en compte la différence de distribution des groupes par rapport au concept mesuré. Elle est dite non conditionnelle si elle examine l'existence de différences dans le fonctionnement d'un item globalement, sans égard à la différence de distribution des groupes comparés (Van de Vijver et Poortinga, 1991, p. 287).

Non conditionnelles par nature, les méthodes basées sur l'analyse de la variance, sur les indices de difficulté ou sur les indices de discrimination (Cardall et Coffman, 1964 ; Green et Draper, 1972 ; Angoff et Ford, 1973 ; Plake et Hoover, 1979-1980 ; Sinnot, 1980 ; Angoff, 1982) exigent relativement peu de sujets, mais elles sont influencées par la différence de distribution des groupes comparés. Ce faisant, elles risquent d'identifier comme biaisés des items qui ne le sont pas, de ne pas détecter des items biaisés ou de donner lieu à des méprises sur la direction du biais (Hunter, 1975 ; Lord, 1977 et 1980 ; Shepard, Camilli et Williams, 1985 ; Camilli et Shepard, 1987 ; Linn et Drasgow, 1987 ; Dorans, 1989 ; Scheuneman et Bleistein, 1989 ; Angoff, 1993). Par ailleurs, leur transformation en méthodes conditionnelles par l'appariement préalable des sujets score à score est une solution dont l'efficacité est limitée. L'appariement peut s'avérer difficile si les groupes comptent peu de sujets et que la différence d'habileté est importante. Restreindre l'étude aux seuls cas où l'appariement est possible entraîne une perte d'information aux extrémités de l'échelle d'habileté.

Les méthodes basées sur les modèles paramétriques ou non paramétriques de réponses aux items, les méthodes basées sur les tableaux de contingence et les méthodes basées sur la régression sont autant de méthodes naturellement conditionnelles. Parce qu'elles prennent en compte la différence de distribution des groupes comparés au regard du concept mesuré, elles risquent moins de confondre une différence réelle d'habileté entre les groupes et une différence due à l'appartenance à un groupe ou à une variable qui lui est associée (Green, 1991, p. 2). Pour cette raison, elles paraissent préférables même si elles nécessitent des échantillons plus nombreux. Les méthodes basées sur les modèles paramétriques à un ou à deux paramètres suscitent cependant des réserves lorsqu'on les applique à des items à choix multiple.

Les modèles paramétriques de réponses aux items fournissent un outil de choix pour estimer la probabilité des réponses en fonction de l'habileté des sujets. La valeur des estimations est cependant liée à la réalisation des postulats sous-jacents, et les méthodes d'analyse du fonctionnement différentiel des items basées sur ces modèles risquent de confondre fonctionnement différentiel des items et mauvais ajustement au modèle. Lorsqu'un test se compose de questions à choix multiple, il y a toujours une possibilité que des sujets très faibles puissent répondre correctement à certains items de façon aléatoire. Étant le seul à prendre ce facteur en considération, le modèle à trois paramètres est le plus approprié. Toutefois, même en utilisant ce modèle, il arrive que des items ne s'ajustent pas au modèle (Lord, 1980 ; Ramsay, 1991 a, p. 612 ; Dorans, 1989, p. 228). Le problème est peu fréquent pour des tests élaborés par des spécialistes dans des conditions rigoureuses. Il peut l'être davantage pour des tests élaborés dans des conditions moins favorables. Il ne peut qu'être plus accentué lorsqu'on lui substitue le modèle à un ou à deux paramètres. Le risque de confondre fonctionnement différentiel des items et mauvais ajustement au modèle en appliquant le modèle à un paramètre à des tests composés d'items à choix multiple est d'ailleurs reconnu et bien étayé (Rudner, Getson et Knight, 1980 b ; Shepard, Camilli et Averill, 1981 ; Scheuneman et Bleistein, 1989 ; Dorans, 1989 ; Angoff, 1993).

Les méthodes du khi carré, la méthode de Mantel-Haenszel et les méthodes basées sur les modèles log-linéaires ou logit reposent sur la répartition des données dans un tableau de contingence à trois dimensions et l'utilisation de tests statistiques. Mis à part les méthodes basées sur les modèles log-linéaires ou logit, elles ne font aucun présupposé sur la forme de la relation entre l'habileté des sujets et la probabilité d'une réponse correcte. De plus, elles peuvent prendre en compte la non monotonicité de la relation, le cas échéant. Ces méthodes diffèrent les

unes des autres par les hypothèses qu'elles permettent de vérifier, la puissance des tests statistiques utilisés pour déceler les items qui ont un fonctionnement différentiel uniforme (FDU) et leur capacité de déceler les items qui ont un fonctionnement différentiel non uniforme (FDNU).

Pour chaque niveau d'habileté considéré sur l'échelle d'habileté, les réponses du groupe focalisé, F, et du groupe de référence, R, sont classées en deux catégories selon que la réponse est correcte ou non. Le tableau ci-dessous illustre la répartition des données et permet de comprendre les formules relatives à la méthode du khi carré complet et à la méthode de Mantel-Haenszel.

Groupes	Réponses à l'item au niveau d'habileté j		
	correctes (1)	incorrectes (2)	Total
R	A_j	B_j	n_{Rj}
F	C_j	D_j	n_{Fj}
Total	m_{1j}	m_{0j}	T_j

Les méthodes basées sur le khi carré sont à la fois les plus simples et les moins puissantes. Elles comprennent la méthode du khi carré de Scheuneman (1979), avec indices orientés ou non orientés, et la méthode du khi carré complet, également avec indices orientés ou non orientés. Seules les valeurs calculées par la méthode du khi carré complet respectent les conditions d'utilisation du khi carré et peuvent se distribuer selon la loi du khi carré (Baker, 1981 ; Camilli, 1979 ; Ironson, 1982, p. 132 ; Mellenbergh, 1982, p. 112 ; Scheuneman et Bleistein, 1989, p. 260). Cette méthode vérifie si la probabilité de répondre correctement à un item est la même à tous les niveaux d'habileté pour chacun des groupes comparés ou si une différence existe à un

niveau quelconque d'habileté. Les hypothèses sont vérifiées à l'aide du khi carré de Pearson avec j(2 - 1) (2 - 1) degrés de liberté, soit

$$\chi_p^2 = \sum_{j=1}^{K} \left[\frac{(A_j - E(A_j))^2}{n_{Rj} n_{Fj} m_{0j} m_{1j} / T^3} \right] , \qquad (1)$$

où

$$E(A_j) = n_{Rj} m_{1j} / T_j . \qquad (2)$$

La méthode permet de déceler simultanément et de façon indifférenciée toute espèce de différence dans le fonctionnement d'un item (Marascuilo et Slaughter, 1981), aussi bien un fonctionnement différentiel uniforme (FDU) qu'un fonctionnement différentiel non uniforme (FDNU).

La méthode de Mantel-Haenszel cherche à accroître la puissance du test statistique en réduisant la portée de l'hypothèse alternative. Elle a été empruntée à la recherche médicale et adaptée à l'étude du fonctionnement différentiel des items par Holland et Thayer (1986 et 1988). Essentiellement, elle vérifie si le rapport entre la probabilité de répondre correctement à un item, P, et la probabilité d'y répondre incorrectement, Q, est le même à tous les niveaux d'habileté pour les deux groupes comparés ou si une différence constante existe entre les groupes. C'est le rapport classique des probabilités qui est vérifié, soit

$$\alpha = P_{Rj} Q_{Fj} / Q_{Rj} P_{Fj} , \qquad (3)$$

pour les j niveaux d'habileté considérés. Le rapport correspond à l'unité s'il n'y a pas de différence entre les groupes. Il a une valeur différente de l'unité s'il existe une différence

constante entre les groupes. L'hypothèse est vérifiée à l'aide du khi carré de Mantel-Haenszel, X^2_{MH}, à un degré de liberté. La formule se présente comme suit :

$$\chi^2_{MH} = \frac{(\left| \sum_j A_j - \sum_j E(A_j) \right| - 1/2)^2}{\sum_j Var(A_j)} \qquad (4)$$

où

$$Var(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)} . \qquad (5)$$

La méthode fournit également une mesure de l'ampleur du fonctionnement différentiel. Celle-ci résulte de l'estimation du rapport classique des probabilités pour tous les niveaux d'habileté,

$$\alpha_{MH} = \frac{\sum A_j D_j / T_j}{\sum B_j C_j / T_j} . \qquad (6)$$

Ce rapport prend des valeurs entre 0 et l'infini avec une valeur égale à l'unité lorsqu'il n'y a pas de différence dans le fonctionnement d'un item. Les valeurs inférieures à l'unité indiquent que la probabilité de donner une réponse correcte est plus grande pour le groupe focalisé ; les valeurs supérieures à l'unité, qu'elle est plus grande pour le groupe de référence. Ce rapport peut être transformé en valeur delta au moyen de la formule

$$\hat{\Delta}_{MH} = -\frac{4}{1,7} ln(\alpha_{MH}) = -2,35 \; ln(\alpha_{MH}). \qquad (7)$$

Selon les critères établis par l'Educational Testing Service (ETS), les valeurs du Δ_{MH} inférieures à l'unité correspondent à un fonctionnement différentiel négligeable. Les valeurs entre 1,00 et 1,50 indiquent un fonctionnement différentiel indésirable sans être très important ; les valeurs supérieures à 1,50, un fonctionnement différentiel important. Outre les Δ_{MH}, on utilise parfois des Z_{MH}. Tout comme les Δ_{MH}, les Z_{MH} sont une transformation linéaire du rapport classique des probabilités estimé. Ils sont donnés par la formule

$$Z_{MH} = \frac{1}{1,7} ln(\alpha_{MH}) . \qquad (8)$$

Les Z_{MH} ont une moyenne de 0 et un écart-type de 1. L'absence de différence dans le fonctionnement d'un item se traduit par une valeur d'indice nulle.

La méthode logit est aussi une tentative d'accroître la puissance du test statistique. Utilisée pour la première fois par Mellenbergh (1982), elle consiste à vérifier à tour de rôle l'existence d'un fonctionnement différentiel non uniforme (FDNU), puis celle d'un fonctionnement différentiel uniforme (FDU). La méthode repose sur les modèles log-linéaires que l'on transforme en modèles logit pour les appliquer à des items notés de façon dichotomique. Elle part du principe que le modèle qui décrit parfaitement les données comporte un effet global, C ; un effet dû au niveau d'habileté, S_i ; un effet dû à l'appartenance à un groupe, G_j ; et un effet dû à l'interaction entre le niveau d'habileté et l'appartenance à un groupe, $(SG)_{ij}$. Le modèle saturé prend la forme

$$Ln(F_{ij1} / F_{ij2}) = C + S_i + G_j + (SG)_{ij} . \qquad (9)$$

Les modèles non saturés utilisés pour vérifier s'il y a fonctionnement différentiel des items sont :

$$Ln(F_{ij1} / F_{ij2}) = C + S_i + G_j \qquad (10)$$

et

$$Ln(F_{ij1} / F_{ij2}) = C + S_i \, . \qquad (11)$$

Dans un premier temps, on vérifie si l'effet d'interaction est nécessaire pour décrire les données, auquel cas on conclut à un fonctionnement différentiel non uniforme (FDNU). Si l'effet d'interaction n'est pas nécessaire, on vérifie la nécessité de l'effet dû à l'appartenance à un groupe. Dans l'affirmative, on conclut à un fonctionnement différentiel uniforme (FDU). Pour vérifier la nécessité d'un effet, on détermine jusqu'à quel point le modèle qui tient compte de cet effet s'ajuste aux données et le modèle qui n'en tient pas compte ne s'y ajuste pas. La vérification est effectuée à l'aide du khi carré de Pearson ou du rapport de vraisemblance G^2. Si le modèle avec effet dû à l'appartenance à un groupe mais sans effet d'interaction s'ajuste aux données et que le modèle avec effet dû au niveau d'habileté mais sans effet dû à l'appartenance à un groupe s'y ajuste également, on prend en compte la différence entre les deux modèles. On conclut à un fonctionnement différentiel uniforme (FDU) si la différence est significative. Dans le cas contraire, on conclut à l'absence de fonctionnement différentiel, l'effet dû au niveau d'habileté étant le seul nécessaire pour décrire les données.

La méthode logit, la méthode de Mantel-Haenszel et les méthodes du khi carré considèrent la variable habileté comme une variable nominale. Cette dernière étant ordonnée et continue, il en résulte une perte d'information. Pour obvier à cette perte, Swaminathan et Rogers (1990)

proposent d'utiliser la régression logistique. L'habileté des sujets et la probabilité d'une réponse correcte sont considérées comme des variables continues. La relation entre les deux variables est décrite par la formule :

$$P(u_{ij} = 1 \mid \theta_{ij}) = \frac{exp(\beta_{0j} + \beta_{1j}\theta_{ij})}{1 + exp(\beta_{0j} + \beta_{1j}\theta_{ij})} \qquad (12)$$

où

$$i = 1,...n \text{ et } j = 1, 2.$$

Dans ce contexte, u_{ij} est la réponse d'une personne i dans le groupe j. P représente la probabilité d'une réponse correcte eu égard aux réponses u des candidats i dans le groupe j et à leur habileté θ_{ij}. β_{0j} représente l'intercept de la courbe de régression, et β_{1j}, la pente de cette même courbe pour le groupe j.

Ce modèle définit les courbes de régression logistique pour chacun des deux groupes comparés. En principe, il n'y a pas de fonctionnement différentiel si les courbes de chaque groupe coïncident. Il y a fonctionnement différentiel uniforme (FDU) si les pentes coïncident et que les intercepts diffèrent. Il y a fonctionnement différentiel non uniforme (FDNU) si les pentes diffèrent, peu importe que les intercepts coïncident ou non. Swaminathan et Rogers (1990) utilisent un test statistique qui vérifie l'existence d'une différence due à l'appartenance à un groupe ou à l'interaction entre cette appartenance et l'habileté des sujets. Ce test a été élaboré à partir de l'équation :

$$P(u = 1) = \frac{e^{z}}{(1 + e^{z})} \qquad (13)$$

où

$$Z = \tau_0 + \tau_1 \theta + \tau_2 G + \tau_3 (\theta G) . \qquad (14)$$

Dans ce cas, τ_1 correspond au coefficient de régression pour l'habileté des sujets, τ_2 correspond au coefficient de régression pour l'appartenance à un groupe ; τ_3, au coefficient de régression pour l'interaction entre le groupe et l'habileté des sujets. Il y a fonctionnement différentiel uniforme (FDU) si τ_2 diffère de 0, et que τ_3 est égal à 0. Il y a fonctionnement différentiel non uniforme (FDNU) si τ_3 diffère de 0, peu importe que τ_2 soit égal à 0 ou qu'il en diffère. Par conséquent, les deux hypothèses à vérifier sont τ_2 est égal à 0 et τ_3 est égal à 0. Swaminathan et Rogers (1990) vérifient les deux hypothèses simultanément avec un test du khi carré à deux degrés de liberté. Ils proposent aussi d'appliquer un test du khi carré qui vérifie séparément l'existence d'une différence au niveau de la pente et de l'intercept. Le nombre de degrés de liberté passe alors de deux à un pour chaque hypothèse vérifiée. Dans la mesure où l'hypothèse d'une relation monotone et continue se réalise, le procédé permet une meilleure utilisation de l'information disponible. Il peut cependant s'avérer moins efficace que la méthode logit ou la méthode de Mantel-Haenszel si l'hypothèse d'une relation monotone et continue ne tient pas.

La méthode du khi carré complet, la méthode de Mantel-Haenszel, la méthode logit et la méthode de régression logistique utilisent le score observé comme variable d'appariement. La méthode de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993, 1995) utilise le score latent. Cette dernière n'est pas, à proprement parler, une méthode d'analyse du fonctionnement différentiel des items, mais bien une méthode d'estimation de la probabilité des réponses en fonction de l'habileté des sujets. Basée sur un modèle non paramétrique de réponses aux items, elle définit une courbe qui décrit la probabilité des réponses en fonction de l'habileté des sujets.

La comparaison à vue des courbes obtenues auprès de deux groupes distincts de sujets permet de déterminer si un item fonctionne différemment d'un groupe à l'autre.

Pour estimer la probabilité des réponses sans estimer de paramètres, la méthode a recours à un procédé de lissage (kernel smoothing) qui exploite le principe de moyenne locale. L'estimation des probabilités d'une réponse m à l'item i, au niveau d'habileté θ, $P_{im}(\theta_q)$, est donnée par la moyenne pondérée des indicateurs de réponse, y_{im}, associés à chaque candidat a, soit :

$$\hat{P}_{im}(\theta_q) = \sum_{a=1}^{N} w_{aq} y_{ima} , \qquad (15)$$

$$= \frac{\sum_{a=1}^{N} K\,[\,(\theta_a - \theta_q) / h\,]\, y_{ima}}{\sum_{a=1}^{N} K\,[\,(\theta_a - \theta_q) / h\,]} . \qquad (16)$$

K représente la fonction de lissage, et h, un paramètre d'étendue locale qui contrôle la quantité de données pondérées. θ_q correspond à la valeur d'habileté estimée au point d'estimation q en appliquant le principe de moyenne locale. Plus spécifiquement, c'est la moyenne des indicateurs pour les valeurs de θ_a qui se situent entre les limites de deux points d'estimation $(\theta_{q-1} + \theta_q)/2$ et $(\theta_q + \theta_{q+1})/2$.

La procédure suivie se divise en quatre étapes : la mise en rang des sujets en se basant sur une estimation de leur habileté ; l'attribution d'un quantile à chaque rang, lequel est utilisé comme estimation de l'habileté des sujets, θ_a ; le regroupement des patrons de réponses en fonction du

rang ; et l'estimation de la probabilité des réponses, $P_{im}(\theta_q)$, par lissage de la courbe qui décrit la relation entre le vecteur d'habilité $\theta_1,....,\theta_N$ et le vecteur des indicateurs binaires de réponse y_{ima}. Le vecteur de réponses a une longueur de N ; le paramètre d'étendue locale, h, une valeur qui s'approche de $N^{-1/5}$. Le quantile attribué à chaque rang découle d'une distribution postulée a priori. Ramsay (1989, 1991a, 1991b, 1992 a, 1992 b, 1993, 1995) postule une distribution normale standard. La mise en rang initiale des sujets repose sur le nombre total de réponses correctes.

La méthode permet d'estimer la probabilité des réponses correctes, mais aussi celle des mauvaises réponses. Le procédé est extrêmement rapide et rend possible la prise en considération des mauvaises réponses pour estimer l'habileté des sujets. Ce faisant, l'estimation de l'habileté des sujets les plus faibles s'avère, en principe, plus précise que l'estimation obtenue en ne considérant que les bonnes réponses. Pour étudier le fonctionnement différentiel des items, Ramsay (1992 b, p. 87) suggère simplement d'analyser les données pour les deux groupes réunis, d'estimer le score latent à partir des probabilités obtenues, puis d'analyser les données de chaque groupe séparément en utilisant les estimations d'habileté obtenues lors de la première analyse. Il suffit ensuite de comparer les courbes de chaque groupe, réunies dans un même graphique, pour voir si l'item affiche un fonctionnement différentiel.

À l'origine, la méthode décrite par Ramsay (1992 b) pour identifier les items qui ont un fonctionnement différentiel ne comportait ni indice pour juger de l'importance du fonctionnement différentiel ni test statistique pour identifier les items qui ont un fonctionnement différentiel. Dorans, Potenza et Ramsay (1993) ont proposé de calculer un indice de standardisation modifié en se servant des probabilités de réponses obtenues pour chaque groupe après l'application du

procédé de lissage. L'estimation de la différence de proportion d'une réponse pour chacun des deux groupes comparés se fait alors aux divers points d'estimation des probabilités de réponses, de manière à couvrir toute l'étendue des scores. Les différences obtenues sont ensuite pondérées par le nombre de sujets du groupe de standardisation, le groupe focalisé ou le groupe de référence, aux divers points d'estimation des probabilités. Comme dans la méthode de standardisation, des différences standardisées supérieures ou égales à 0,05 ou à 0,10 et des différences standardisées inférieures à -0,05 ou -0,10, selon le critère retenu, pourraient devenir l'indice d'un fonctionnement différentiel.

Ramsay a aussi envisagé la possibilité d'utiliser les probabilités obtenues après l'application du procédé de lissage pour calculer un rapport de probabilité similaire au rapport α_{MH} et de le transformer en un Δ_{MH} modifié. Dans ce cas,

$$\Delta_m = -2,35\, ln(\alpha) \qquad (17)$$

où

$$\alpha_m = \sum \frac{N_A P_1(\theta)\, N_D Q_2(\theta)}{N_C P_2(\theta)\, N_B Q_1(\theta)} . \qquad (18)$$

N représente alors le nombre de personnes qui ont répondu correctement (N_A et N_C) ou incorrectement (N_B et N_D) dans le groupe de référence (N_A et N_B) et le groupe focalisé (N_C et N_D). P_1 représente la probabilité d'une réponse dans le groupe de référence ; P_2, la probabilité d'une réponse dans le groupe focalisé ; Q_1, la différence entre la probabilité d'une réponse et l'unité dans le groupe de référence et Q_2, la différence correspondante dans le groupe focalisé. θ

représente l'habileté des sujets sur l'échelle d'habileté. L'identification des items qui ont un fonctionnement différentiel repose alors sur l'ampleur des indices. Elle pourrait se faire à partir des critères utilisés par l'Educational Testing Service (ETS) avec la méthode de Mantel-Haenszel (voir à la page 11).

Finalement, Ramsay a intégré le calcul d'indices modifiés de standardisation au programme qu'il a créé pour l'application de son modèle (Ramsay, 1993 et 1995). La formule utilisée s'énonce comme suit :

$$\beta_F = \sum_{(q=1)}^{Q} pr_q \left[P_{im}^{(F)}(\theta) - P_{im}^{(R)}(\theta) \right] \qquad (19)$$

Dans ce cas, pr_q correspond à la proportion de personnes dans le groupe de référence qui affichent un niveau d'habileté θ_q. $P^{(R)}{}_{im}(\theta)$ correspond à la probabilité des réponses en fonction de l'habileté des sujets dans le groupe de référence. $P^{(F)}{}_{im}(\theta)$ renvoie à la probabilité correspondante dans le groupe focalisé. β_F est ici une différence de difficulté standardisée modifiée. La méthode utilisée pour obtenir ces différences de difficulté devrait donner des indices plus stables que la méthode de standardisation régulière avec des échantillons de taille réduite. C'est du moins l'avantage escompté. Il n'est pas dit, toutefois, que les autres méthodes applicables à de petits échantillons ne soient pas plus stables ni plus efficaces.

Même si la méthode de standardisation n'a pas été conçue pour s'appliquer à de petits échantillons, il nous apparaît utile d'en donner un aperçu, en raison des efforts qui ont été faits pour la rendre applicable à des échantillons de taille réduite. Comme la méthode de Mantel-

Haenszel ou les méthodes du khi carré, elle ne fait aucun présupposé quant à la forme de la relation entre les réponses à l'item et l'habileté des sujets. Cette absence de présupposé permet de tenir compte de la non monotonicité de la relation et des écarts possibles par rapport à un modèle paramétrique imposé d'emblée. En soi, la méthode est essentiellement descriptive. Sa validité tient à l'utilisation de grands échantillons. Pour cette raison, elle permet l'étude du fonctionnement différentiel des items dans la population. Son application à de petits échantillons risque cependant de donner des indices exagérément gonflés (Schmitt, 1987). Mise au point par Dorans et Kulick (1983), elle a donné lieu à de nombreuses recherches (Dorans, Schmitt et Bleistein, 1988 ; Dorans, Schmitt et Bleistein, 1992 ; Rivera et Schmitt, 1988 ; Schmitt, 1988 ; Schmitt et Bleistein, 1987 ; Schmitt et Dorans, 1990 ; Schmitt, Dorans, Crone et Maneckshava, 1991). Dorans et Kulick (1986) et, plus tard, Dorans (1989), puis Dorans et Holland (1993) en font la description détaillée.

En gros, la méthode consiste à déterminer la proportion de réponses correctes pour chaque groupe comparé à chacun des niveaux possibles d'habileté au regard du concept mesuré. Les proportions ainsi obtenues sont conditionnelles. Elles permettent de tracer une courbe empirique et non paramétrique de la probabilité des réponses en fonction de l'habileté des sujets. Les courbes de chaque groupe donnent un aperçu visuel des différences dans le fonctionnement des items et de leur ampleur. Un diagramme de dispersion montre avec plus de précision les différences dans le fonctionnement des items et la grandeur de ces différences. Outre ces représentations graphiques, la méthode fournit deux indices statistiques : la différence de difficulté standardisée, DDS, et la racine du carré des différences de difficulté pondérées, RCDDP.

La différence de difficulté standardisée est de loin l'indice le plus utilisé. Celle-ci se définit comme suit :

$$DDS = \sum_{s=1}^{S} K_s(P_{fs} - P_{rs}) \; / \; \sum_{s=1}^{S} K_s \; . \qquad (20)$$

$K_s / \sum K_s$ représente un facteur de pondération utilisé à chaque niveau d'habileté s pour pondérer les proportions de réponses correctes du groupe focalisé P_{fs} et du groupe de référence P_{rs}. Dépendant du nombre de sujets dans le groupe de standardisation, ce facteur permet de dissocier impact et fonctionnement différentiel des items. Le groupe utilisé est déterminé par le chercheur, de manière à assurer les résultats les plus stables. Plusieurs options sont possibles :

$K_s = N_{ts}$,	le nombre de personnes dans le groupe total dont l'habileté est s ;
$K_s = N_{rs}$,	le nombre de personnes dans le groupe de référence dont l'habileté est s ;
$K_s = N_{fs}$,	le nombre de personnes dans le groupe focalisé dont l'habileté est s ;
$K_s = N_n$,	le nombre relatif dans un groupe de référence constituant la norme.

En général, on utilise le groupe focalisé, soit $K_s = N_{fs}$, parce que celui-ci confère un poids plus grand aux différences de difficulté associées aux niveaux d'habileté les plus souvent atteints par le groupe focalisé.

La différence de difficulté standardisée, DDS, peut prendre des valeurs qui varient entre -1,00 et 1,00. Les valeurs positives indiquent que l'item favorise le groupe focalisé ; les valeurs négatives, qu'il favorise le groupe de référence. Les valeurs comprises entre -0,05 et 0,05 sont considérées comme négligeables. Les valeurs comprises entre -0,05 et -0,10 et celles comprises entre 0,05 et 0,10 deviennent l'indice d'un fonctionnement différentiel qui, sans être très important, mérite un examen plus poussé. Les valeurs supérieures à 0,10 ou inférieures à -0,10

sont considérées comme l'indice d'un fonctionnement différentiel important. Selon Dorans (1986 p. 226), les items auraient alors un fonctionnement différentiel facilement explicable.

La racine du carré des différences de difficulté pondérées, RCDDP, a été mise au point pour pallier le fait que les différences positives de difficulté sont annulées par les différences négatives. Le but était de permettre la détection des items qui ont un fonctionnement différentiel non uniforme (FDNU). Selon Wright (1886 et 1987), ce nouvel indice,

$$RCDDP = [\sum_{s=1}^{S} K_s(P_{fs} - P_{rs})^2 / \sum_{s=1}^{S} K_s]^{0.5}, \qquad (21)$$

est biaisé, parce qu'il retient toutes les erreurs d'échantillonnage à chaque niveau d'habileté. Pour cette raison, on n'utilise plus que la différence de difficulté standardisée, DDS.

État de la recherche

Plusieurs recherches ont été entreprises pour vérifier la validité et la fidélité des méthodes conditionnelles d'analyse du fonctionnement différentiel des items. Ces recherches utilisent soit des données réelles, soit des données simulées. Les recherches sur la validité étudient la convergence des méthodes et leur capacité de détecter des items qui sont réputés avoir un fonctionnement différentiel. Dans le second cas, on a recours à un fonctionnement différentiel simulé, à des items choisis pour avantager un des groupes comparés ou encore on privilégie une méthode pour identifier les items avec un fonctionnement différentiel.

Les premières recherches sur la convergence des méthodes mettent en cause des méthodes basées sur des indices de difficulté transformés (IDT), sur le khi carré ou sur les modèles paramétriques de réponses aux items à un ou à trois paramètres (Ironson et Subkoviak, 1979 ; Merz et Grossen, 1979 ; Rudner, Getson et Knight, 1980 a ; Shepard, Camilli et Averill, 1981 ; Shepard, Camilli et Williams, 1985). Tant les recherches avec des données réelles (Ironson et Subkoviak, 1979 ; Shepard, Camilli et Averill, 1981 ; Shepard, Camilli et Williams, 1985) que les recherches avec des données simulées (Merz et Grossen, 1979 ; Rudner, Getson et Knight, 1980 a ; Shepard, Camilli et Williams, 1985) font état de relations moyennes entre les méthodes basées sur le khi carré et des méthodes basées sur le modèle à trois paramètres. Se situant plus souvent qu'autrement entre 0,50 et 0,65, ces relations sont généralement supérieures à celles observées pour des méthodes basées sur les indices de difficulté transformés. Compte tenu des relations observées, plusieurs chercheurs considèrent les méthodes basées sur le khi carré comme un substitut valable aux méthodes basées sur le modèle à trois paramètres (Ironson, 1982 ; Shepard, Camilli et Williams, 1985 ; Crocker et Algina, 1986 ; Seong et Subkoviak, 1987). Si les

relations observées sont suffisantes pour justifier l'emploi des méthodes du khi carré au lieu des méthodes basées sur le modèle à trois paramètres, la détection des items avec un fonctionnement différentiel demeure problématique.

Quelques recherches ont également été effectuées pour démontrer la validité de la méthode logit (Mellenbergh, 1982 ; Van der Flier, Mellenbergh, Ader et Wijn, 1984 ; Kok, Mellenbergh et Van der Flier, 1985). Aucune ne compare cette méthode à une méthode basée sur la théorie de réponses aux items et le modèle à trois paramètres. Parmi les recherches effectuées, l'une utilise des données réelles obtenues après avoir soumis deux groupes de sujets à des items qui devaient normalement les favoriser (Kok, Mellenbergh et Van der Flier, 1985). L'autre utilise des données simulées (Van der Flier, Mellenbergh, Ader et Wijn, 1984). Dans les deux cas, le taux de détection des items qui sont réputés avoir un fonctionnement différentiel montre une capacité réelle de détecter des items avec un fonctionnement différentiel. Comme avec la méthode du khi carré ou les méthodes basées sur les modèles paramétriques de réponses aux items, des items avec fonctionnement différentiel peuvent ne pas être détectés et des items sans fonctionnement différentiel peuvent être identifiés à tort comme ayant un fonctionnement différentiel.

Au milieu des années quatre-vingt, l'attention s'est portée sur la méthode de Mantel-Haenszel. Plusieurs recherches ont été réalisées. Les unes la comparent à d'autres méthodes (Skaggs et Lissitz, 1988 ; Hambleton, Rogers et Arrasmith, 1988 ; Baghi et Ferrara, 1989, 1990 ; Hambleton et Rogers, 1988 et 1989 ; Engelhard, Anderson et Gabrielson, 1990 ; Shermis et St-George, 1990 ; Sykes et Fitzpatrick, 1990 ; Hambleton et Jones, 1992 ; Ibrahim, 1992 ; Linacre et Wright, 1988 ; Raju, Drasgow et Slinde, 1993 ; Schulz, Perlman, Rice et Wright, 1989). Les autres vérifient les facteurs susceptibles d'avoir un effet sur la détection des items qui ont un

fonctionnement différentiel (Camilli et Smith, 1990 ; De Mauro, 1990 ; Zwick et Ercikan, 1989 ; Clauser, Mazor et Hambleton, 1991 a, 1991 b et 1991 c ; Mazor, Clauser et Hambleton, 1991 a et 1991 b ; Donohue, Holland et Thayer, 1993 ; Miller et Oshima, 1992 ; Spray 1989). La plupart des recherches sur la convergence des méthodes utilisent des données réelles (Skaggs et Lissitz, 1988 ; Baghi et Ferrara, 1989, 1990 ; Hambleton et Rogers, 1988 et 1989 ; Engelhard, Anderson et Gabrielson, 1990 ; Hambleton et Jones, 1992 ; Raju, Drasgow et Slinde, 1993 ; Schulz, Perlman, Rice et Wright, 1989). Inversement, la plupart des recherches axées sur la capacité de détecter les items qui ont un fonctionnement utilisent des données simulées (Clauser, Mazor et Hambleton, 1991 a, 1991 b et 1991 c ; Mazor, Clauser et Hambleton, 1991 a et 1991 b ; Donohue, Holland et Thayer, 1993 ; Miller et Oshima, 1992).

En se basant sur le pourcentage d'accord, la convergence observée avec des méthodes basées sur la théorie des réponses aux items (TRI) et le modèle à trois paramètres (Skaggs et Lissitz, 1988 ; Hambleton, Rogers et Arrasmith, 1988 ; Baghi et Ferrara, 1990 ; Hambleton et Rogers, 1988 et 1989) est passable pourvu que le nombre de sujets soit assez grand. Bien que la méthode de Mantel-Haenszel repose sur des assises différentes, Holland et Thayer (1986 et 1988) ont établi un lien entre cette méthode et le modèle à un paramètre. Plusieurs recherches ont vérifié la justesse de ce lien. Ici encore, la convergence observée est acceptable (Baghi et Ferrara, 1989 ; Engelhard, Anderson et Gabrielson, 1990 ; Sykes et Fitzpatrick, 1990 ; Schulz, Perlman, Rice et Wright, 1993). Par ailleurs, les recherches qui reposent sur la capacité de déceler les items qui ont un fonctionnement différentiel simulé (Clauser, Mazor et Hambleton, 1991 a et 1991 b ; Mazor, Clauser et Hambleton, 1991 a et 1991 b ; Donohue, Holland et Thayer, 1993 ; Ibrahim, 1992 ; Miller et Oshima, 1992) confirment sa capacité de détecter des items qui ont un fonctionnement différentiel. Toutefois, tous les items ne sont pas nécessairement détectés et,

comme nous le verrons plus loin, plusieurs facteurs peuvent avoir un effet sur la détection des items qui ont un fonctionnement différentiel.

Plus récentes que les recherches sur la méthode de Mantel-Haenszel, les recherches sur la méthode de régression logistique sont aussi moins nombreuses. Les premières recherches remontent au début des années quatre-vingt-dix (Rogers et Swaminathan, 1990 ; Swaminathan et Rogers, 1990). Ces dernières utilisent des données simulées. La comparaison des taux de détection montre que le procédé est, à toutes fins utiles, aussi efficace que la méthode de Mantel-Haenszel pour déceler les items qui ont un fonctionnement différentiel uniforme (FDU). Il s'avère nettement plus efficace pour les items qui ont un fonctionnement différentiel non uniforme (FDNU). D'autres recherches tenteront de préciser cette tendance (Brown, 1992 ; Ibrahim, 1992 ; Ochieng, 1992 ; Pang et Boss, 1993 ; Pang, Tian et Boss, 1994: Tian, Pang et Boss, 1994 a et 1994 b). Mis à part l'étude effectuée par Ibrahim (1992), toutes se limitent à une comparaison avec la méthode de Mantel-Haenszel. Certaines ont recours à des données simulées (Ibrahim, 1992 ; Ochieng, 1992 ; Pang et Boss, 1993). D'autres font appel à des données réelles (Brown, 1992 ; Pang, Tian et Boss, 1994 ; Tian, Pang et Boss, 1994 a et 1994 b). Toutes révèlent une capacité de détecter des items qui ont un fonctionnement différentiel. Reste à déterminer jusqu'à quel point la méthode peut le faire et les facteurs qui influent sur sa capacité de détecter des items qui ont un fonctionnement différentiel.

Quant à la méthode de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995), la recherche en est encore à l'état embryonnaire. En se servant de données réelles, Dorans, Potenza et Ramsay (1993) et Liu, Dorans et Ramsay (1995) ont examiné la possibilité d'appliquer la méthode de standardisation modifiée par l'application du procédé de Ramsay. L'avantage réside

dans le lissage des courbes de réponses et l'utilisation du score latent comme variable d'appariement. Liu, Dorans et Ramsay (1995) ont montré que le procédé peut donner des indices proches des différences de difficulté standardisée régulière (DDS). Toutefois, la distribution des indices diffère. Si le fonctionnement différentiel est positif, les valeurs absolues des différences de difficulté standardisée régulière (DDS) sont plus petites. Si le fonctionnement différentiel est négatif, elles sont plus grandes. Les différences de difficulté observées sont minimes, ce qui restreint la portée de l'étude. Dorans, Potenza et Ramsay (1993) ont également examiné la variabilité des indices d'un échantillon à un autre de même taille pour des échantillons de taille variée. Leur étude révèle que la moyenne des indices pour 100 échantillons de même taille n'est pas influencée par la taille des groupes comparés. Par contre, les écarts-types le sont. Des échantillons de 200 sujets par groupe s'avèrent trop petits pour produire des estimations stables des différences de difficulté standardisée modifiée (DDSM). Par contre des échantillons de 800 sujets par groupe offrent des indices relativement stables ; des échantillons de 3200 sujets par groupe, des indices très stables. L'étude porte uniquement sur six items : trois qui ont un fonctionnement différentiel très important (une DDSM plus grande ou égale à 0,15 en valeur absolue) et trois qui ont un fonctionnement différentiel négligeable (une DDSM plus petite ou égale à 0,02 en valeur absolue). Étant donné la particularité des données, d'autres études sont nécessaires avant de pouvoir tirer des conclusions définitives.

Par ailleurs, si la capacité de détecter des items qui ont un fonctionnement différentiel à l'aide de la méthode du khi carré complet, de la méthode de Mantel-Haenszel, de la méthode logit et de la méthode de régression logistique est acquise, la généralisabilité des résultats n'est pas assurée. Plusieurs recherches ont été effectuées pour vérifier la fidélité de ces méthodes (Doolittle, 1983 ; Hoover et Kolen, 1984 ; Loyd, 1984 ; Perlman, Bezruezko, Junker, Reynolds, Rice et Schulz,

1988 ; Hambleton et Rogers, 1989 ; Engelhard, Anderson et Gabrielson, 1990 ; Ryan, 1990 et 1991). Toutes utilisent des données réelles, une décision pleinement justifiée, puisqu'on s'intéresse à l'effet des fluctuations d'échantillonnage. Règle générale, on y examine la stabilité des indices entre deux ou plusieurs échantillons pris deux à deux et celle des décisions prises à partir de ces mêmes indices quant à la présence ou à l'absence de fonctionnement différentiel. Une seule étude porte sur la cohérence interne des décisions, quels que soient les échantillons (Perlman, Bezruezko, Junker, Reynolds, Rice et Schulz, 1988). La fidélité est donnée par la proportion de variance vraie sur la variance totale pour 30 échantillons de même taille. La variance vraie correspond à la variance à l'intérieur des échantillons ; la variance d'erreur, à la variance entre les échantillons.

Les recherches sur la stabilité des indices obtenus par la méthode du khi carré complet, la méthode logit ou la méthode de Mantel-Haenszel indiquent une stabilité qui varie de faible à modérée pour des échantillons de 500 sujets par groupe ou moins (Doolittle, 1983 ; Hoover et Kolen, 1984 ; Loyd, 1984 ; Hambleton et Rogers, 1989 ; Raju, Bode et Larsen, 1988 et 1989 ; Engelhard, Anderson et Gabrielson, 1990 ; Ryan, 1990 et 1991). La stabilité est toujours moindre pour les indices qualitatifs (le classement des items en items qui ont ou qui n'ont pas de fonctionnement différentiel, à partir des indices quantitatifs et d'un critère de décision) que pour les indices quantitatifs (une mesure du fonctionnement différentiel des items telle la valeur calculée du khi carré, le rapport classique des probabilités ou sa transformation en valeur delta). Le choix de la méthode ne semble pas avoir d'effet, puisqu'on observe la même tendance pour la méthode du khi carré complet (Doolittle, 1983 ; Hoover et Kolen, 1984), la méthode de Mantel-Haenszel (Hambleton et Rogers, 1989 ; Engelhard, Anderson et Gabrielson, 1990 ; Ryan, 1990 et 1991) et la méthode logit (Loyd, 1984). Pour ce qui est de la cohérence des décisions, l'étude de

Perlman, Bezruezko, Junker, Reynolds, Rice et Schulz (1988) montre que la méthode du khi carré de Mantel-Haenszel n'atteint une fidélité acceptable que si le nombre total de sujets est d'au moins 666 sujets par groupe. La fidélité varie alors de 0,64 à 0,77. La méthode n'atteindrait le seuil de 0,80 que si les échantillons comptent au moins 1000 sujets par groupe. D'autres études seraient nécessaires pour confirmer la justesse de cette tendance. Toutefois, des études qui portent sur la détection des items qui ont un fonctionnement différentiel pointent dans la même direction.

Plusieurs facteurs peuvent influer sur la capacité de détecter des items qui ont un fonctionnement différentiel et la généralisabilité des résultats d'une analyse du fonctionnement différentiel des items. Certains sont liés à des caractéristiques inhérentes aux méthodes ; d'autres, à des caractéristiques inhérentes aux groupes comparés ou aux items analysés. Un facteur souvent négligé dans l'interprétation des résultats des études comparatives est la capacité de déceler les items qui ont un fonctionnement différentiel non uniforme (FDNU). Contrairement à la méthode logit ou à la méthode de régression logistique, la méthode de Mantel-Haenszel n'est pas conçue pour déceler les items qui ont un fonctionnement différentiel non uniforme (FDNU). Il arrive qu'elle puisse déceler certains items de ce genre lorsque le groupe qui réussit le mieux change à une des extrémités de l'échelle d'habileté (Rogers et Swaminathan, 1990 ; Mazor, Clauser et Hambleton, 1991 a). Son inaptitude à déceler les items qui ont un fonctionnement différentiel non uniforme (FDNU) surviendrait lorsque le groupe qui réussit le mieux change au centre de la distribution d'habileté (Gutierrez, 1989 ; Ibrahim, 1992 ; Mazor, Clauser et Hambleton, 1991 a). S'appuyant sur ce principe, Mazor, Clauser et Hambleton (1991 a) proposent de reprendre les analyses de fonctionnement différentiel après avoir divisé les groupes comparés en deux, selon que les sujets obtiennent un score au-dessus ou au-dessous de la moyenne. À l'aide de données

simulées, ils montrent que le procédé permet de détecter une proportion importante d'items avec fonctionnement différentiel non uniforme (FDNU). Ces items ne sont pas détectés lorsqu'on utilise l'ensemble des sujets, et le procédé n'augmente pas indûment l'erreur de type I. Ceci dit, Hambleton et Rogers (1989) notent que des modifications moins importantes de l'étendue de la distribution des scores sur l'échelle d'habileté peuvent faire varier le nombre d'items avec un fonctionnement différentiel non uniforme (FDNU) détectés. Par ailleurs, les items avec un fonctionnement différentiel non uniforme (FDNU) pourraient être source d'instabilité dans les indices (Loyd, 1984). À notre connaissance, aucune recherche systématique n'a été effectuée pour confirmer la justesse de cette hypothèse.

Mis à part la capacité d'identifier les items qui ont un fonctionnement différentiel non uniforme (FDNU), les principaux facteurs susceptibles d'avoir un effet sur l'identification des items qui ont un fonctionnement différentiel concernent le regroupement des sujets en niveaux d'habileté plus ou moins étendus, la taille des groupes comparés, leur différence d'habileté et leur plus ou moins grande disproportion, le degré de difficulté des items, leur discrimination et l'ampleur des différences dans le fonctionnement de l'item. L'effet dû au regroupement des sujets en niveaux d'habileté qui réunissent plusieurs scores possibles ne concerne que les méthodes basées sur des tableaux de contingence. Selon Ironson (1982), de trois à cinq niveaux d'habileté suffiraient pour contrôler la différence d'habileté des groupes avec les méthodes du khi carré. Loyd (1983) a vérifié la justesse de cette assertion avec la méthode du khi carré complet. L'étude révèle un nombre nettement plus grand d'items avec un fonctionnement différentiel lorsqu'il n'y a que trois niveaux d'habileté, c'est-à-dire lorsque le risque de confondre une différence réelle d'habileté entre les groupes et une différence due à un biais est le plus grand.

Les autres études qui portent sur le regroupement des données concernent la méthode de Mantel-Haenszel. Raju, Bode et Larsen (1989) ont démontré à l'aide de données réelles que la corrélation entre les indices de fonctionnement différentiel des items issus de cette méthode s'approche d'une corrélation parfaite peu importe le nombre de niveaux considérés (2, 4, 6, 8 ou 10 pour un test de 40 items). Leur étude révèle également un nombre toujours plus grand d'items avec un fonctionnement différentiel lorsqu'il n'y a que deux niveaux d'habileté. Lorsqu'il y a quatre niveaux d'habileté ou plus, le pourcentage d'accord entre deux analyses qui utilisent un nombre différent de niveaux d'habileté est nettement plus grand. Par ailleurs, le nombre d'items avec un fonctionnement différentiel varie peu quel que soit le nombre de niveaux d'habileté. Il serait cependant prématuré de conclure qu'il suffit de quatre niveaux d'habileté pour contrôler la différence d'habileté des groupes comparés sans nuire à la justesse des résultats. Le nombre optimal d'intervalles de scores nécessaires pour assurer des résultats stables pourrait être lié à la longueur du test critère et à l'étendue des scores possibles. Wright (1986 et 1987) a noté que six niveaux d'habileté donnaient des résultats moins stables que 61 pour une variable d'appariement qui compte 600 scores possibles.

Plus récemment, Clauser, Mazor et Hambleton (1991 a) ont utilisé des données simulées pour vérifier l'influence du nombre de niveaux d'habileté sur l'identification des items avec un fonctionnement différentiel à l'aide de la méthode de Mantel-Haenszel. Pour ce faire, ils ont postulé des tests de 80 items et réparti les données en 2, 5, 10, 20 et 81 niveaux d'habileté pour des échantillons de 100, de 200, de 500, de 1000 et de 2000 sujets par groupe. Leur étude prend en compte la différence d'habileté des groupes et révèle un effet différent. Pour des groupes de même habileté en regard de la variable d'appariement, le nombre d'items qui ont un fonctionnement différentiel et qui sont dépistés augmente en fonction du nombre de sujets. Par

contre, il varie peu en fonction du nombre de niveaux. Par ailleurs, le nombre d'items faussement identifiés comme ayant un fonctionnement différentiel est à peu près nul. Pour des groupes d'habileté inégale, le nombre d'items dépistés croît également avec le nombre de sujets, mais on note aussi une tendance à avoir nettement plus d'items avec un fonctionnement différentiel lorsqu'il y a moins de niveaux. De plus, la proportion d'items faussement identifiés comme ayant un fonctionnement différentiel est plus grande lorsqu'il y a moins de niveaux, et d'autant plus grande qu'il y a plus de sujets dans les groupes. Clauser, Mazor et Hambleton (1991a) en concluent que le regroupement des sujets en niveaux d'habileté réunissant plusieurs scores ne devrait pas entraîner une réduction importante du nombre maximal de niveaux possibles. À leur avis, l'augmentation du nombre de sujets est préférable au regroupement des données pour accroître la capacité de détecter des items qui ont un fonctionnement différentiel. Les résultats obtenus avec la méthode de Mantel-Haenszel sont susceptibles de se généraliser à toute méthode basée sur le regroupement de données dans un tableau de contingence. L'ampleur de l'effet peut cependant varier selon les méthodes.

D'autres études confirment l'influence de la taille des échantillons et de la différence de distribution des groupes sur le taux de détection des items qui ont un fonctionnement différentiel et laissent présager une influence de la disproportion des groupes. Comparant la méthode de Mantel-Haenszel et la méthode de régression logistique à l'aide de données simulées, Rogers et Swaminathan (1990) observent un taux de détection plus grand des items qui ont un fonctionnement différentiel lorsqu'ils utilisent des échantillons de 500 sujets par groupe au lieu d'échantillons de 250 sujets par groupe, et ce, quelle que soit la méthode et le type de fonctionnement différentiel des items. Dans une étude conçue pour mettre en valeur la supériorité de la méthode de régression logistique sur la méthode de Mantel-Haenszel,

Swaminathan et Rogers (1990 a et 1990 b) rapportent également une augmentation du taux de détection avec l'augmentation du nombre de sujets pour des groupes de 250 et de 500 sujets par groupe.

Utilisant des données réelles au lieu de données simulées pour comparer la méthode de Mantel-Haenszel et la méthode de régression logistique, Brown (1992) observe une tendance similaire pour les deux méthodes. Le taux de détection des items qui ont un fonctionnement différentiel croît avec l'augmentation de la grandeur des groupes pour des échantillons de 100, de 200, de 300, de 500, de 750 et de 1000 sujets par groupe. Comparant la méthode de Mantel-Haenszel à diverses autres méthodes en se servant de données réelles, Baghi et Ferrera (1989) ont également noté un effet de la taille des groupes sur le khi carré et l'alpha de Mantel-Haenszel ainsi que sur les taux de détection basés sur le khi carré de Mantel-Haenszel. Le test analysé compte 45 items et les sujets sont groupés en 4, 5, 6, 7, 8 et 10 niveaux d'habileté pour le calcul du khi carré. Dans l'étude de Brown, on considère autant de niveaux d'habileté qu'il y a de scores possibles.

Dans le but d'avoir une meilleure compréhension de la méthode de Mantel-Haenszel, Mazor, Clauser et Hambleton (1991 b) ont également étudié l'influence de la taille des échantillons sur le taux de détection des items qui ont un fonctionnement différentiel. Leur étude prend en compte la différence d'habileté des groupes comparés. Utilisant des données simulées, les chercheurs ont constitué des tests de 75 items et créé des données pour des échantillons de 2000, de 1000, de 500, de 200 et de 100 sujets par groupe. Ayant recours à un procédé en deux étapes comme Clauser, Mazor et Hambleton (1991 a), ils considèrent autant de niveaux d'habileté qu'il y a de scores possibles. Comme dans l'étude de Clauser, Mazor et Hambleton (1991 a), on note un taux de détection qui augmente avec le nombre de sujets et un taux de détection plus grand pour les

groupes de même habileté. On note cependant que l'augmentation du taux de détection n'est pas proportionnelle à l'augmentation du nombre de sujets. Sauf exception, les items qui ont un fonctionnement différentiel dans les petits échantillons ont également un fonctionnement différentiel dans les grands. Plus les échantillons sont petits, plus la différence de difficulté des items identifiés comme ayant un fonctionnement différentiel est grande. La différence de difficulté des items détectés tend à être plus grande lorsque les groupes comparés ne sont pas de même habileté et que leur distribution en regard de la variable d'appariement diffère. Il y a aussi plus d'items faussement identifiés comme ayant un fonctionnement différentiel, mais leur nombre demeure limité.

L'étude de Mazor, Clauser et Hambleton (1991 b) utilise le khi carré de Mantel-Haenszel pour déterminer si un item a un fonctionnement différentiel ou non. L'influence de la taille des échantillons sur le taux de détection des items qui ont un fonctionnement différentiel est prévisible dans ce contexte, l'augmentation du nombre de sujets étant un moyen d'accroître la puissance d'un test statistique. Elle est susceptible de se produire peu importe la méthode lorsqu'on utilise un test de signification pour identifier les items potentiellement biaisés. L'ampleur de l'effet peut cependant varier selon les méthodes. Quoique moins évidente, l'influence de la différence d'habileté et de l'inégalité des groupes comparés est également prévisible. Les modifications de fréquences engendrées par la disproportion des groupes et la différence de distribution en regard de la variable d'appariement entraînent une modification des fréquences théoriques utilisées pour le calcul du khi carré et, par ricochet, une modification de la valeur calculée du khi carré. La question est plutôt de savoir si la taille des groupes comparés influe sur les indices de fonctionnement différentiel basés sur l'ampleur du fonctionnement différentiel et quelle méthode est la plus valide et la plus fiable sous ce rapport.

Dans une étude qui porte exclusivement sur la méthode de Mantel-Haenszel, Gutierrez (1989) a démontré à l'aide de données simulées que la taille des groupes comparés avait un effet sur la distribution des Z_{MH}. L'écart-type des Z_{MH} varie en fonction de la taille des groupes comparés, de la difficulté des items et de leur discrimination. L'augmentation du nombre de sujets entraîne une diminution de la grandeur de l'écart-type. Celle-ci augmente avec l'accroissement de la discrimination. Lorsque les items sont de difficulté moyenne, la grandeur de l'écart-type diminue. La diminution est d'autant plus grande que les échantillons comptent plus de sujets. La différence est d'autant plus prononcée que la discrimination est plus importante. La valeur des indices aux percentiles $P_{2,5}$, P_5, P_{95} et $P_{97,5}$ tend aussi à décroître avec l'augmentation de la taille des groupes comparés et elle augmente avec la discrimination.

Dans une étude qui met en cause la méthode de Mantel-Haenszel et la méthode de régression logistique avec test de signification distinct pour les items qui ont un fonctionnement différentiel uniforme (FDU) et les items qui ont un fonctionnement différentiel non uniforme (FDNU), Ochieng (1992) et Ibrahim (1992) constatent que la grandeur des échantillons n'a pas d'effet significatif sur les indices de fonctionnement différentiel obtenus en appliquant la méthode de régression logistique pour items avec fonctionnement différentiel uniforme (FDU). Elle a cependant un effet significatif sur la distribution des X^2_{MH} et des Δ_{MH}, ainsi que sur les indices résultant de l'application de la méthode de régression logistique pour items avec fonctionnement différentiel non uniforme (FDNU). Les deux études ont été effectuées à l'aide de données simulées et de groupes disproportionnés. Pang et Boss (1993) ont effectué une étude similaire avec des données simulées et des groupes égaux. Il en résulte que la taille des groupes comparés n'aurait pas d'effet sur les indices de fonctionnement différentiel obtenus en appliquant la méthode de régression logistique, et ce, que les items aient un fonctionnement différentiel

uniforme (FDU) ou un fonctionnement différentiel non uniforme (FDNU). Par contre, la valeur critique des statistiques aux percentiles P_{90} et P_{95} est légèrement surestimée lorsque les items ont un fonctionnement différentiel uniforme (FDU) ou qu'ils ont un fonctionnement différentiel non uniforme (FDNU) et une discrimination élevée. Cette surestimation peut faire en sorte que des items qui n'ont pas de fonctionnement différentiel soient identifiés comme en ayant un.

Dans le même ordre d'idée, Tian, Pang et Boss (1994 a et 1994 b) et Pang, Tian et Boss (1994) ont vérifié l'effet dû à la grandeur des échantillons sur la méthode de régression logistique et la méthode de Mantel-Haenszel, en se servant de données réelles. L'intérêt de ces études réside dans l'utilisation des valeurs Z_{MH} comme indice de fonctionnement différentiel au lieu de la valeur calculée du khi carré de Mantel-Haenszel. Pang, Tian et Boss (1994) étudient le fonctionnement différentiel des items pour des groupes égaux. Des échantillons de 100, de 250, de 500 et de 1000 sujets par groupes ont été sélectionnés de façon aléatoire parmi les hommes et les femmes qui ont subi un test de mathématiques de l'American College Testing. La sélection a été effectuée de façon à avoir 30 groupes distincts de 100, de 250, de 500 et de 1000 sujets par groupe. Comme on utilise des données réelles, on ne connaît pas les items qui ont un fonctionnement différentiel. Pour obvier au problème, on a classé les items en quatre catégories selon que les items n'ont pas de fonctionnement différentiel ou qu'ils ont un fonctionnement différentiel certain, probable ou possible. La classification se fonde sur le Z_{MH} moyen obtenu pour les 30 échantillons de taille maximale. Un item est considéré avoir un fonctionnement différentiel certain si le Z_{MH} moyen est plus grand ou égal à 0,25 en valeur absolue. Il est considéré avoir un fonctionnement différentiel probable si le Z_{MH} moyen se situe entre 0,20 et 0,25, abstraction faite de l'orientation des différences dans le fonctionnement des items. Enfin, il est réputé avoir un fonctionnement différentiel possible s'il se situe entre 0,15 et 0,20 en valeur

absolue. D'après cette étude, la méthode de régression logistique est légèrement plus puissante que la méthode de Mantel-Haenszel pour détecter les items qui ont fonctionnement différentiel uniforme. Elle présente cependant un taux d'items faussement identifiés comme ayant un fonctionnement différentiel plus grand que la méthode de Mantel-Haenszel. Quelle que soit la méthode, une augmentation de la taille des échantillons tend à accroître le taux d'items faussement identifiés comme ayant un fonctionnement différentiel. Dans l'ensemble, les deux méthodes s'avèrent d'une efficacité comparable pour détecter les items qui ont un fonctionnement différentiel uniforme (FDU). Seule la méthode de régression logistique peut déceler les items qui ont un fonctionnement différentiel non uniforme (FDNU). La généralisabilité des résultats d'un échantillon à un autre de même taille ou de taille supérieure est moins assurée qu'elle ne l'est pour les items qui ont un fonctionnement différentiel uniforme (FDU). Peu importe la méthode, des échantillons de 500 sujets par groupe sont nécessaires pour déceler les items qui ont un fonctionnement différentiel uniforme (FDU) certain avec un minimum de fiabilité. Des échantillons de 1000 sujets par groupe sont nécessaires dans le cas d'items dont le fonctionnement différentiel uniforme (FDU) est probable ou possible.

En utilisant une approche méthodologique similaire à celle de Pang, Tian et Boss (1994), Tian, Pang et Boss (1994 a et 1994 b) obtiennent des résultats qui vont dans le même sens. Dans les deux cas, la taille des groupes comparés exerce une forte influence sur le taux de détection des items qui ont un fonctionnement différentiel. Celui-ci est d'autant plus grand qu'il y a plus de sujets dans les groupes comparés. Des échantillons d'au moins 500 personnes par groupe sont nécessaires pour déceler les items qui ont un fonctionnement différentiel certain et des échantillons de 1000 sujets par groupe pour les items qui ont un fonctionnement différentiel probable ou possible. Tian, Pang et Boss (1994 b) notent que la moyenne des Z_{MH} est

relativement stable quelle que soit la taille des échantillons, mais l'écart-type diminue quand la taille des groupes comparés augmente. Dans l'ensemble, la méthode de Mantel-Haenszel s'avère moins sujette à des variations lorsque le nombre de sujets est plus grand. Elle peut déceler des items qui ont un fonctionnement différentiel non uniforme lorsque les items sont très difficiles, non lorsque les items sont de difficulté moyenne. Dans cette dernière éventualité, les courbes de réponses correctes se croisent au centre de la distribution.

Résumé de la situation

Les recherches sur la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995) sont encourageantes, mais insuffisantes pour tirer des conclusions sur la validité et la fidélité de l'approche. Les recherches sur la méthode de Mantel-Haenszel et la méthode de régression logistique révèlent une capacité réelle de détecter des items qui ont un fonctionnement différentiel. Toutefois, des items avec un fonctionnement différentiel peuvent ne pas être détectés et des items sans fonctionnement différentiel peuvent être identifiés comme en ayant un. Non seulement le nombre total d'items détectés peut varier d'un échantillon à un autre de même taille, mais les items identifiés comme ayant un fonctionnement différentiel peuvent varier. Par ailleurs, la taille des groupes comparés peut avoir un effet sur la précision des indices et la détection des items qui ont un fonctionnement différentiel.

Dans une perspective purement pragmatique, une question se pose. Quelle méthode et quelle grandeur d'échantillon sont les plus appropriées pour détecter des items qui ont un fonctionnement différentiel tout en s'assurant que les résultats des analyses effectuées sont fiables et valides ? Pour répondre à cette question, il faut considérer d'une part la stabilité des indices et celle des décisions prises à partir de ces mêmes indices, d'autre part, le taux de détection que l'on peut espérer pour les items qui ont un fonctionnement différentiel eu égard à la taille des échantillons et à la grandeur des différences dans le fonctionnement des items.

Les recherches sur la stabilité des indices et celle des décisions prises à partir de ces derniers ne concernent que la méthode de Mantel-Haenszel (Perlman, Bezruezko, Junker, Reynolds, Rice et

Schulz, 1988 ; Engelhard, Anderson et Gabrielson, 1990 ; Ryan, 1990 et 1991). Mis à part l'étude de Perlman et de ses collègues en 1988, ces recherches n'ont pas pour but de déterminer le nombre de sujets requis dans les échantillons pour s'assurer d'indices stables et de décisions fiables. Par ailleurs, l'étude de Perlman et de ses collègues repose sur des données réelles. Par conséquent, les résultats sont tributaires du test analysé et des personnes à qui il a été administré. Pour cette raison, des recherches avec d'autres tests et d'autres échantillons sont nécessaires pour confirmer les tendances observées.

Plusieurs recherches examinent l'effet de la taille des groupes comparés sur la distribution des indices fournis par la méthode de Mantel-Haenszel et la méthode de régression logistique (Gutierrez, 1989 ; Ochieng, 1992 ; Ibrahim, 1992 ; Pang et Boss, 1993) ou sur leur capacité de détecter des items qui ont un fonctionnement différentiel (Swaminathan et Rogers, 1990 b ; Ibrahim, 1992 ; Mazor, Clauser et Hambleton, 1991 b ; Brown, 1992 ; Tian, Pang et Boss, 1994 a et 1994 b ; Pang, Tian et Boss, 1994). Les recherches qui considèrent l'effet de la taille des groupes comparés sur la distribution des indices fournis par la méthode de Mantel-Haenszel ou la méthode de régression logistique (Gutierrez, 1989 ; Ochieng, 1992 ; Ibrahim, 1992 ; Pang et Boss, 1993) montrent un effet possible de la taille des groupes comparés. Toutefois, ces recherches ne sont pas conçues pour déterminer la grandeur d'échantillon souhaitable pour obtenir des résultats fiables. Qui plus est, les résultats ne concordent pas en ce qui concerne la méthode de régression logistique et toutes, à part l'étude de Pang et Boss (1994), reposent sur des groupes inégaux.

Les recherches qui examinent l'effet de la taille des groupes sur la détection des items qui ont un fonctionnement différentiel (Swaminathan et Rogers, 1990 b ; Ibrahim, 1992 ; Mazor, Clauser et

Hambleton, 1991 b ; Brown, 1992 ; Tian, Pang et Boss, 1994 a et 1994 b ; Pang, Tian et Boss, 1994) fournissent une indication de la grandeur d'échantillon souhaitable, mais aucune ne considère la méthode des différences de difficulté standardisée modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995). Dans quelques recherches réalisées avec des données simulées (Swaminathan et Rogers, 1990 b ; Ibrahim, 1992 ; Mazor, Clauser et Hambleton, 1991 b), on a répété les analyses avec plusieurs échantillons de même taille. Le procédé donne un aperçu du taux de détection des items qui ont un fonctionnement différentiel que l'on peut espérer eu égard à la grandeur des groupes comparés. Toutefois, ces recherches ont été conçues pour étudier l'effet de certaines variables. De ce fait, elles ne considèrent que le nombre d'items détectés parmi la totalité des items qui ont un fonctionnement différentiel. De plus, les tests simulés et les caractéristiques des items avec un fonctionnement différentiel manquent souvent de réalisme et suscitent des doutes quant à l'applicabilité des taux dans une situation réelle.

Swaminathan et Rogers (1990 b) postulent que 20 pour cent des items d'un test ont un fonctionnement différentiel, 10 pour cent un fonctionnement différentiel uniforme (FDU) et 10 pour cent, un fonctionnement différentiel non uniforme (FDNU). La proportion d'items avec un fonctionnement différentiel et leur répartition en fonctionnement différentiel uniforme ou fonctionnement différentiel non uniforme ne se fondent sur aucune observation. Ibrahim (1992) postule des tests avec une répartition uniforme et équidistante du paramètre de difficulté et il ne considère qu'un ou deux paramètres de discrimination. De plus, les items avec fonctionnement différentiel se distribuent à des intervalles égaux sur l'échelle d'habileté. Aucun test ne présente de telles caractéristiques dans une situation réelle. En général, il y a plus d'items de difficulté moyenne que d'items très faciles ou très difficiles. On peut donc s'attendre à avoir plus d'items

avec un fonctionnement différentiel parmi les items de difficulté moyenne, puisqu'il y a plus d'items de ce genre. La même remarque vaut pour l'étude de Mazor, Clauser et Hambleton (1990 b). On postule cinq tests de 75 items : 59 items sans fonctionnement différentiel et 16 items avec un fonctionnement différentiel. Dans ce cas, les items avec un fonctionnement différentiel peuvent avoir quatre grandeurs pour le paramètre de discrimination (0,25 ; 0,60 ; 0,90 ou 1,25) et quatre différences de fonctionnement au niveau du paramètre de difficulté (0,25 ; 0,50 ; 1,00 ; 1,50). Par ailleurs, on considère cinq paramètres de difficulté possibles (-2,5 ; -1,0 ; 0,0 ; 1,0 ; 2,5), mais il n'y a que quatre conditions par test. Il en va de même pour l'étude de Clauser, Mazor et Hambleton (1991 a), une autre étude où l'on fournit des taux de détection en fonction de la grandeur des échantillons.

Au manque de réalisme des tests simulés s'ajoute l'impossibilité, pour le lecteur, de déterminer avec précision l'ampleur des différences de fonctionnement simulées. Swaminathan et Rogers (1990 b) expriment celle-ci par une différence de surface. Ibrahim (1992) et Mazor, Clauser et Hambleton (1991 b) l'expriment par une différence au niveau du paramètre de difficulté ou du paramètre de discrimination. Quelle que soit l'étude, on ne connaît pas la répercussion des différences simulées sur la probabilité d'une réponse correcte, ni au niveau du paramètre de difficulté, ni pour l'ensemble de l'échelle d'habileté. Mazor, Clauser et Hambleton (1990b) présentent quatre figures : deux qui illustrent des cas extrêmes et deux, des cas intermédiaires. Des deux cas extrêmes illustrés, l'un présente une différence de fonctionnement d'un ou deux pour cent au niveau de la probabilité d'une réponse correcte (paramètre de discrimination de 0,25 et différence du paramètre de difficulté de 0,25). La différence est, à toutes fins utiles, négligeable. L'autre illustre une différence si grande qu'il est à peu près impossible qu'on ne puisse la détecter, peu importe la méthode (paramètre de

discrimination de 1,25 et différence du paramètre de difficulté de 1,50). De fait, la différence maximale au niveau de la probabilité d'une réponse correcte est de l'ordre de 55 pour cent environ. Dans une étude où ils utilisent les mêmes paramètres et les mêmes différences de difficulté pour simuler des items avec un fonctionnement différentiel, Clauser, Mazor et Hambleton (1991 b) stipulent que les différences de fonctionnement simulées se traduisent par une différence dans la probabilité d'une réponse correcte qui varie de 0,01 à 0,52 selon les items. Le taux de détection que l'on peut espérer pour les items qui ont un fonctionnement différentiel est susceptible de varier en fonction de l'ampleur des différences dans le fonctionnement des items. Que peuvent valoir les taux observés si les items qui ont un fonctionnement différentiel simulé présentent des différences infimes ou exagérément grandes ? Quelle est l'utilité des taux de détection si les différences de fonctionnement simulées ne sont pas interprétables en terme de différence dans la probabilité des réponses ?

Enfin, il faut souligner que Swaminathan et Rogers (1990 b) ne considèrent que deux grandeurs d'échantillon (250 et 500 sujets par groupe). Qui plus est, les analyses ne sont répétées que pour les échantillons de 500 sujets par groupe et il n'y a que 20 répétitions. Mazor, Clauser et Hambleton (1991 b) ne font pas d'analyses répétées. Ils simulent cinq tests avec 16 items ayant un fonctionnement différentiel dans chaque test et ils évaluent les résultats pour les 80 items avec un fonctionnement différentiel simulé. Seul Ibrahim (1992) procède avec 100 répétitions, un minimum pour que l'erreur-type maximale possible des taux de détection ne dépasse pas 0,05.

Quelques recherches sur la capacité de détecter des items qui ont un fonctionnement différentiel en fonction de la taille des groupes comparés utilisent des données réelles (Brown, 1992 ; Tian, Pang et Boss, 1994 a et 1994 b ; Pang, Tian et Boss, 1994). Ces recherches ont l'avantage de

fournir des taux de détection qui reposent sur des données réalistes. Par contre, parce qu'on utilise des données réelles, le besoin de reprendre les études avec d'autres tests subsiste pour confirmer les tendances observées. Nonobstant ce trait commun à toute recherche effectuée avec des données réelles, les recherches réalisées suscitent des réserves.

Tout d'abord, il convient de souligner que Brown (1992), Tian, Pang et Boss (1994 a et 1994 b) et Pang, Tian et Boss (1994) examinent la capacité de détecter des items réputés avoir un fonctionnement différentiel, uniquement dans le cas où l'on utilise le khi carré de Mantel-Haenszel ou les khis carrés pour la régression logistique. Ce qui arriverait si l'on devait utiliser le delta de Mantel-Haenszel ou les différences de difficulté standardisée modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995) demeure à explorer. Par ailleurs, Brown (1992) n'utilise que cinq répétitions au maximum et les cinq répétitions ne s'appliquent pas à toutes les grandeurs d'échantillon examinées. Qui plus est, la détermination des items réputés avoir un fonctionnement différentiel repose sur des critères circonstanciels qui varient selon la taille des échantillons et les groupes comparés. Enfin, les taux de détection rapportés ne tiennent pas compte de l'ampleur des différences dans le fonctionnement des items.

Les études réalisées par Tian, Pang et Boss (1994 a et 1994 b) et Pang, Tian et Boss (1994) ont l'avantage d'utiliser 30 répétitions, ce qui représente une nette amélioration par rapport à l'étude de Brown (1992), puisque l'erreur-type maximale possible passe alors de 0,224 à 0,091 pour le pourcentage d'items avec un fonctionnement différentiel détectés. Qui plus est, la méthode utilisée pour déterminer quels items ont un fonctionnement différentiel et faire état de la proportion d'items détectés pour des échantillons de même taille prend en compte la grandeur

moyenne des différences de fonctionnement de chaque item pour l'ensemble de l'échelle d'habileté. Pour ce faire, on utilise le Z_{MH} , et les items sont ensuite répartis en trois catégories selon la grandeur du Z_{MH}. Ce dernier est une transformation linéaire du rapport classique moyen des probabilités, l'alpha de Mantel-Haenszel. Quoique des plus valables, cette mesure est difficile à évaluer en terme de différence dans la probabilité d'une réponse correcte. De plus, pour établir le taux de détection que l'on peut espérer pour les items qui sont réputés avoir un fonctionnement différentiel, il serait souhaitable de s'appuyer sur plus de 30 répétitions.

CHAPITRE III

BUT DE L'ÉTUDE

Nous avons vu au chapitre précédent que la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995) avait été très peu étudiée. L'apport de ce modèle n'est pas sans intérêt. Le procédé permet l'utilisation du score latent au lieu du score observé pour estimer l'habileté des sujets. Il prend en considération les mauvaises réponses pour estimer le score latent. Ce faisant, il fournit une estimation plus précise de l'habileté des sujets que celle obtenue en ne considérant que les bonnes réponses. Le gain se fait particulièrement sentir chez les sujets dont l'habileté est plus faible. L'application du procédé de lissage atténue l'effet des cas extrêmes sur l'estimation de la probabilité des réponses en fonction de l'habileté des sujets et, par ricochet, sur les indices de fonctionnement différentiel qui en découlent. Si les différences de difficulté standardisée ne sont pas appropriées avec des échantillons de taille réduite, les différences de difficulté standardisée modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995) pourraient s'avérer au moins aussi fiables et aussi efficaces que la méthode du delta de Mantel-Haenszel sinon davantage. S'il en était ainsi, le procédé aurait l'avantage de permettre l'étude simultanée du fonctionnement différentiel des bonnes réponses et l'étude du fonctionnement différentiel des leurres, puisque le procédé fournit tous ces indices simultanément. Compte tenu des possibilités offertes, il nous apparaît utile de pousser plus avant l'étude du fonctionnement différentiel des items avec la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995).

Nous avons également vu au chapitre précédent que les recherches effectuées ne permettaient pas de déterminer laquelle de la méthode de Mantel-Haenszel ou de la méthode de régression logistique est la plus appropriée pour s'assurer de résultats fiables. Non seulement ne peut-on pas répondre à cette question, mais on ne peut déterminer de façon satisfaisante l'efficacité relative des méthodes et la grandeur d'échantillon souhaitable pour s'assurer d'une efficacité acceptable. Selon Cohen (1988), la puissance d'un test statistique dépend de trois paramètres : le seuil de signification, la fidélité des statistiques utilisées ou plus exactement leur précision, et la grandeur de l'effet que l'on espère pouvoir détecter dans la population. Même si d'autres facteurs peuvent intervenir, la précision des statistiques utilisées dépend toujours du nombre de sujets. Zimmerman, Williams et Zumbo (1993) ont montré que c'est la réduction de la variance qui accroît l'efficacité d'un test statistique. Comme la grandeur de la variance dépend du nombre de sujets, on peut penser que la taille des échantillons aura un effet sur les méthodes d'analyse du fonctionnement différentiel qui utilisent une mesure de l'ampleur des différences comme il en a sur celles qui utilisent un test statistique. Si l'effet de la taille des groupes comparés sur la fidélité des méthodes d'analyse du fonctionnement différentiel et leur efficacité est reconnu, peu de recherches prennent en compte la grandeur des différences que l'on espère pouvoir détecter dans la population. De fait, seules les études effectuées par Tian, Pang et Boss (1994 a et 1994 b) et Pang, Tian et Boss (1994) le font. Ces recherches utilisent le Z_{MH} moyen pour 30 échantillons de 1000 sujets par groupe. Notre intention est d'étudier l'efficacité des méthodes en fonction de la grandeur des différences que l'on espère pouvoir détecter dans la population et d'utiliser les différences de difficulté standardisée comme mesure de l'ampleur des différences dans le fonctionnement des items. Le Z_{MH} est une transformation du delta de Mantel-Haenszel. Il existe une relation étroite entre le delta de Mantel-Haenszel et les différences de difficulté standardisée. Il existe même une formule pour passer de l'un à l'autre. Toutefois, les différences

de difficulté standardisée sont plus faciles à interpréter en terme de différence dans la probabilité des réponses.

La présente étude a donc pour but d'examiner la fidélité et l'efficacité de la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995). Elle a également pour but de voir comment cette méthode se compare à la méthode du delta ou du khi carré de Mantel-Haenszel et à la méthode de régression logistique. Ce faisant, nous tenterons de déterminer quelle méthode est la plus fiable et la plus efficace et quelle grandeur d'échantillon est souhaitable pour s'assurer de résultats d'une fiabilité et d'une efficacité acceptables. Les questions auxquelles nous nous proposons de répondre se formulent ainsi :

1. Quelle est l'efficacité de la méthode de standardisation modifiée pour détecter les items qui ont un fonctionnement différentiel et la fidélité des indices eu égard à la taille des groupes comparés ?

2. Est-ce que la taille des groupes comparés a un effet sur l'efficacité de la méthode de standardisation modifiée et la fidélité des indices ?

3. Dans quelle mesure l'efficacité de la méthode de standardisation modifiée et la fidélité des indices se comparent à l'efficacité de la méthode de Mantel-Haenszel ou de la méthode de régression logistique et à la fidélité de leurs indices ?

4. De combien faut-il accroître la taille des groupes comparés pour qu'une méthode s'avère aussi efficace qu'une autre de prime abord plus efficace ?

5. Est-ce que l'effet dû à la taille des groupes comparés varie d'une méthode à l'autre de façon importante ?

Pour atteindre les buts que nous nous sommes fixés et répondre aux questions de recherche, nous nous appuyons sur un ensemble de principes, de postulats et d'hypothèses qu'il nous paraît utile préciser pour bien situer l'étude et en dégager l'originalité.

Tout d'abord, il convient de signaler que l'existence d'un fonctionnement différentiel ne correspond pas nécessairement à un biais. Ackerman (1992) établit une distinction entre les habiletés que l'on veut mesurer et les habiletés que l'on ne veut pas mesurer mais qui peuvent jouer un rôle dans la réussite d'un item. Les premières constituent les habiletés valides ; les secondes, des habiletés nuisibles. Si un test est unidimensionnel, un biais correspond à une différence dans les habiletés nuisibles chez des personnes de groupes distincts dont l'habileté valide est identique. L'impact résulte d'une différence entre les groupes au niveau de l'habileté valide. Or, c'est l'interaction entre les sujets et les items qui doit être unidimensionnel. Vu sous cet angle, un test peut être unidimensionnel de trois façons. Les items font appel à plusieurs habiletés, mais les sujets diffèrent sous une seule habileté ou un même composite. Les items ne mesurent qu'une seule habileté ou un même composite de sorte que les différences qui peuvent exister entre les sujets sous d'autres aspects demeurent sans conséquence. Enfin, le test est constitué d'un seul item. Si tous les items mesurent exactement le même composite (habileté valide et habiletés nuisibles), le test sera unidimensionnel. Plus le composite mesuré par chaque item diffère d'un item à l'autre, plus le test est multidimensionnel. Si l'on utilise une méthode d'analyse du fonctionnement différentiel des items qui postule l'unidimensionnalité et que le test est multidimensionnel, on pourra détecter des items qui ont réellement un biais, mais aussi des items non biaisés, la différence de fonctionnement étant alors le reflet d'une différence au niveau

de l'habileté vraie. Il appartient au chercheur de s'assurer que le test est unidimensionnel et que les items sont également valides pour chacun des groupes comparés. Or, il n'y a pas de méthode infaillible pour vérifier l'unidimensionnalité d'un test et s'assurer que tous les items sont également valides pour tous les groupes comparés. La difficulté d'interpréter les items détectés comme des items biaisés lorsqu'on utilise des méthodes d'analyse du fonctionnement différentiel qui postulent l'unidimensionnalité est un fait reconnu. Angoff (1988) écrira à ce propos que les méthodes d'analyse du fonctionnement différentiel des items sont peut-être plus utiles à des fins diagnostiques qu'elles ne le sont pour détecter des items biaisés. La distinction entre biais et impact aurait moins d'importance. À notre avis, pouvoir interpréter les items détectés comme la résultante d'un biais relève de la validité expérimentale. La présente étude porte sur les méthodes d'analyse du fonctionnement différentiel, non sur l'interprétation des items détectés comme des items biaisés.

Nonobstant le but des analyses de fonctionnement différentiel et l'interprétation que l'on peut faire des résultats, l'utilité des méthodes d'analyse du fonctionnement différentiel des items dépend de leur fidélité, c'est-à-dire leur capacité de classer les items de la même façon et d'identifier les mêmes items quels que soient les échantillons d'une même population. Sans fidélité, il n'y a pas de généralisation possible. On pourra effectuer des analyses de fonctionnement différentiel, mais les résultats observés se limiteront à l'échantillon. Si un test est unidimensionnel et que tous les items sont également valides pour chacun des groupes comparés, le fait de ne pas détecter les mêmes items d'un échantillon à un autre tiré d'une même population tient à des fluctuations d'échantillonnage. Si un test est multidimensionnel et qu'on utilise une méthode unidimensionnelle d'analyse du fonctionnement différentiel des items, le score total ne reflétera pas correctement l'habileté des sujets, puisque leur habileté peut varier

selon les dimensions qui expliquent les réponses aux items. Toutefois, la structure multidimensionnelle du test ne devrait pas varier beaucoup d'un échantillon à un autre. Autrement dit, si les items détectés ne sont pas tous le reflet d'un biais, il n'en demeure pas moins que les variations dans les items détectés seront également imputables à des fluctuations d'échantillonnage.

Peu importe la raison d'être des analyses de fonctionnement différentiel, leur utilité est également liée à leur efficacité, c'est-à-dire leur capacité de détecter tous les items qui accusent une différence de fonctionnement d'une grandeur prédéterminée dans la population. Lorsqu'on compare deux groupes distincts de personnes, des différences sont à prévoir. Ces différences peuvent être plus ou moins grandes. Seules les différences qui ont une grandeur jugée pertinente ou suffisamment importante pour retenir l'attention intéressent le chercheur. En principe, une différence pertinente dans le fonctionnement des items de groupes distincts de personne devrait dépasser les différences à l'intérieur des groupes. Au-delà de cette exigence, la grandeur des différences est encore à déterminer.

Que l'on utilise des méthodes d'analyse du fonctionnement différentiel des items basées sur un test statistique ou des méthodes basées sur une mesure de l'ampleur des différences dans le fonctionnement des items, il est postulé que l'efficacité des méthodes dépend du critère utilisé, de la grandeur des différences de fonctionnement que l'on espère pouvoir détecter dans la population et de la précision des statistiques utilisées pour évaluer les différences de fonctionnement des items. La méthode des différences de difficulté standardisée par l'application du modèle de Ramsay, la méthode du khi carré ou du delta de Mantel-Haenszel et la méthode de régression logistique dépendent en dernière analyse des proportions de réponses

correctes à chaque niveau d'habileté. Le nombre de sujets contenus dans un échantillon influe sur la précision de ces proportions et, par ricochet sur celle des différences de proportions observées entre deux groupes. Pour cette raison, il est à prévoir que la taille des groupes comparés influera sur la fidélité et l'efficace de ces méthodes. Reste à déterminer l'importance de l'effet et la grandeur d'échantillon souhaitable.

Toutes les méthodes d'analyse du fonctionnement différentiel des items sont susceptibles de détecter des items qui affichent des différences importantes de fonctionnement. Elles sont davantage susceptibles de se différencier lorsqu'il s'agit de détecter des items dont les différences de fonctionnement ne sont pas très grandes. Pour déterminer quelle méthode est la plus efficace, il y aurait donc lieu de voir comment les méthodes se comportent lorsque les différences de fonctionnement ne sont pas très importantes. Par ailleurs, l'efficacité des méthodes devrait être étudiée dans des conditions aussi réalistes que possible. Comme nous n'avons que très peu d'indications sur la grandeur des différences de fonctionnement auxquelles on peut s'attendre dans une population, il nous apparaît nécessaire d'utiliser des données réelles.

CHAPITRE IV

MÉTHODOLOGIE

Données analysées

La provenance

La présente étude repose sur des données réelles. Celles-ci proviennent d'une banque de données de l'American College Testing Program (ACTP), plus précisément des résultats à la formule 39B de l'épreuve intitulée "The ACT Assessment Test" (AAT). Employée comme test d'admission dans un grand nombre de collèges et d'universités américaines, cette épreuve se compose de quatre tests : un test de mathématiques, un test d'anglais, un test de lecture et un test de sciences sociales. Seules les données du test de mathématiques sont utilisées. Le fonctionnement différentiel des items (FDI) est étudié en fonction du sexe des personnes évaluées. Les personnes de sexe masculin forment le groupe de référence ; les personnes de sexe féminin, le groupe focalisé.

Le test analysé

Le test analysé compte 60 questions à choix multiple. Chaque question présente un choix de cinq réponses. Les questions sont réparties en trois domaines : 24 questions sur l'algèbre fondamentale, 18 questions sur l'algèbre ou la géométrie analytique et 18 questions sur la géométrie plane ou la trigonométrie. La durée du test est de 60 minutes. Les personnes qui subissent l'épreuve sont invitées à répondre au plus grand nombre possible de questions, quitte à

passer les questions qui leur demandent trop de temps et à y revenir s'il leur reste du temps. En soi, le test consiste en un test de rendement (achievement test) de type traditionnel. Les items de chaque domaine s'entremêlent et sont répartis du plus facile au plus difficile. Chaque item consiste en une équation ou un problème à résoudre. Les items nécessitent des connaissances en arithmétique, en algèbre, en géométrie analytique, en géométrie plane ou en trigonométrie, mais les habiletés de raisonnement mathématiques requises pour répondre aux questions demeurent peu complexes pour qui possède les connaissances. Il est postulé que le test est unidimensionnel, hypothèse qui sera évaluée un peu plus loin.

La population cible

Au total, 183 356 personnes de race blanche ont subi le test de mathématiques du "The ACT Assessment Test". Pour la présente étude, la population ciblée est constituée de toutes les personnes de race blanche qui en sont à leur douzième année d'études, qui prévoient obtenir leur diplôme d'études secondaires en 1990 et dont l'anglais est la langue d'usage dominante à la maison. La population ainsi définie compte 160 161 personnes et représente 87,350 pour cent de toutes les personnes de race blanche qui ont subi "The ACT Assessment Test" en 1989. Restreindre ainsi la population initiale s'explique par un souci d'homogénéité. Le fonctionnement différentiel des items doit s'expliquer par une relation entre l'appartenance à un groupe particulier de personnes et des caractéristiques inhérentes aux items ou aux conditions d'administration du test. Des différences au niveau du nombre d'années de scolarité, de la race ou de la langue d'usage dominante peuvent s'associer au sexe et atténuer, amplifier ou expliquer les différences dans le fonctionnement des items.

La sélection d'un échantillon

Pour décrire les caractéristiques du test et vérifier si le test peut être considéré comme unidimensionnel, nous avons sélectionné un échantillon de 40 000 sujets. Notre intention première était d'utiliser toute la population ciblée, mais l'entreprise s'avérait impossible. Pour mener cette étude à terme, il fallait pourvoir travailler sur micro-ordinateur. Le fichier de données aurait pris trop d'espace disque. La grandeur de l'échantillon a été fixée de façon à maintenir l'erreur-type des statistiques qui sont pertinentes à l'étude et qui dépendent du nombre de sujets (moyenne, écart-type, fréquences et proportions de réponses correctes à chaque niveau d'habileté) à un niveau acceptable. La sélection des sujets a été effectuée de façon aléatoire. L'échantillon sélectionné compte 16 521 personnes de sexe masculin et 23 479 personnes de sexe féminin. Les hommes représentent 41,300 pour cent des sujets ; les femmes, 58,700 pour cent des sujets. L'échantillon réunit 24,975 pour cent des personnes de la population ciblée. Cet échantillon sera désormais considéré comme la population d'origine.

Les caractéristiques du test pour la population cible

Règle générale, les hommes obtiennent des résultats globalement plus élevés que les femmes, et la distribution des scores est plus étendue. Dans les deux cas, la distribution des scores est asymétrique, et l'asymétrie est positive. La cohérence interne du test est aussi plus grande pour les hommes que pour les femmes. Le tableau 1 reproduit la moyenne et l'écart-type de chaque groupe pris séparément et la cohérence interne du test pour chacun des deux groupes. En se basant sur le rapport critique, la moyenne et l'écart-type de chaque groupe diffèrent l'un de l'autre de façon significative au seuil de 0,01. Par contre, les coefficients de cohérence interne ne

diffèrent pas l'un de l'autre de façon significative. La figure 2, à l'annexe A, reproduit la distribution des scores de chaque groupe.

Tableau 1. Résultats au test et cohérence interne du test pour les hommes et les femmes pris séparément et pour les deux groupes réunis.

Test de mathématiques (N = 60 items)				
Groupe	**Nombre de sujets**	**Moyenne**	**Écart-type**	**Alpha de Cronback**
Hommes	16 521	28,565	13,988	0,919
Femmes	23 479	25,162	12,661	0,896
Hommes et femmes	40 000	26,567	13,209	0,909

L'évaluation de l'unidimensionnalité

Il a été postulé que le test analysé était unidimensionnel. Il n'existe pas de tests parfaitement et rigoureusement unidimensionnels, et l'unidimensionnalité recherchée est relative (Hambleton et Swaminathan, 1985 ; Hulin, Drasgow et Parsons, 1983). Un test est considéré comme unidimensionnel si un trait domine nettement les autres et que ce trait explique une partie importante des résultats. Il n'y a pas de méthode ni de critère qui fassent l'unanimité parmi les chercheurs pour démontrer l'existence d'un trait prédominant lorsque les données sont dichotomiques. Plusieurs chercheurs ont fait des suggestions et des recommandations. Lord (1980) suggère le recours à l'analyse en composantes principales pour un premier aperçu. La suggestion a été reprise par de nombreux chercheurs (Hattie, 1985 ; Hambleton et Swaminathan, 1985 ; Hulin, Drasgow et Parsons, 1983). Bien qu'elle prête à critique, l'utilisation de l'analyse

en composantes principales est devenue pratique courante pour qui veut élaborer un test à partir de modèles unidimensionnels de réponses aux items. Son emploi repose sur l'existence d'un lien entre le modèle de l'ogive normale et le modèle de l'analyse en composantes principales, lien qui a été démontré par Lord et Novick (1968) et expliqué par Hulin, Drasgow et Parsons (1983).

Dans cette perspective, Reckase (1979) soutient qu'un test peut être considéré comme unidimensionnel s'il satisfait aux deux conditions suivantes. L'analyse en composantes principales révèle une première composante qui explique au moins 20 pour cent de la variance totale des scores. La valeur de la première racine latente est nettement plus grande que la valeur de la deuxième racine latente. On considère généralement que la première racine latente doit être au moins quatre fois plus grande que la deuxième pour satisfaire à la deuxième condition. Non préoccupés de satisfaire aux exigences de modèles unidimensionnels de réponses aux items, Collins, Cliff, McCormick et Zatkin (1986) proposent l'utilisation du rapport L pour évaluer la grandeur de l'écart entre les racines latentes et déterminer le nombre optimum de composantes à extraire. Ce rapport correspond à la différence entre les racines latentes k et k + 1 sur la différence entre les racines latentes k + 1 et k + 2, pour toute racine latente k et k + 1 dont la différence est supérieure à 0,100. On retient les k premières racines latentes pour lesquelles le rapport L est quatre fois plus grand que le rapport L subséquent. Laforge (1981) suggère l'emploi du test de sphéricité de Bartlett pour déterminer le nombre optimum de racines à extraire. Enfin, Stout (1987) propose un test statistique qui s'appuie sur le principe d'indépendance locale pour vérifier si un test est assez unidimensionnel pour permettre l'application d'un modèle unidimensionnel de réponses aux items ou pour déterminer la structure dimensionnelle d'un test.

Pour évaluer l'unidimensionnalité du test analysé, nous avons effectué des analyses en composantes principales pour les deux groupes réunis ainsi que pour les deux groupes pris séparément. Les analyses ont été effectuées à partir de coefficients de corrélation phi et de coefficients de corrélation tétrachorique. Les deux types de coefficients conduisent aux mêmes conclusions. Toutefois, les corrélations, la grandeur des racines latentes et la proportion de variance expliquée par le premier facteur sont moins grandes lorsqu'on utilise les corrélations phi. Après rotation des facteurs, les items semblent vouloir se grouper en fonction de la difficulté statistique des items. Ceci dit, les pondérations initiales, avant rotation, présentent des caractéristiques très semblables à celles obtenues pour les coefficients de corrélation tétrachorique. L'examen des figures qui illustrent la relation entre la grandeur des racines latentes et l'ordre des facteurs montre un premier facteur nettement prédominant et un deuxième facteur qui se situe au point d'inflexion de la courbe sur laquelle s'alignent les valeurs des racines latentes. De telles figures montrent que le test analysé peut être considéré comme unidimensionnel (voir les figures 3 et 4 à l'annexe A).

L'analyse en composantes principales réalisée à partir des coefficients de corrélation tétrachorique montre pour les deux groupes réunis une première composante qui explique 25,351 pour cent de la variance totale. Cette première composante est près de neuf fois plus grande que la deuxième, et le rapport L est 20 fois plus grand. Qui plus est, on trouve des items de tous les domaines sous le premier facteur, et, quel que soit le domaine, les items peuvent avoir des pondérations élevées. On trouve également des items de tous les domaines avec des pondérations importantes sous le deuxième facteur. Bref, on ne peut établir de lien entre le domaine de contenu auquel appartiennent les items et les composantes (voir les tableaux 1, 2, et 3 à l'annexe A). D'ailleurs, Pang, Tian et Boss (1994) ont utilisé le score total au test et le score

aux items du domaine pour étudier le fonctionnement différentiel des items du test pour les hommes et les femmes de race blanche qui ont subi l'épreuve. Leur étude révèle peu de différence dans les items détectés lorsqu'ils utilisent les scores de chaque domaine au lieu du score total.

Par contre, le deuxième facteur pourrait s'expliquer par la limite de temps imposée pour répondre aux questions, laquelle serait source d'erreur dans la mesure. L'hypothèse est soutenue par un taux d'abstention qui croît au fur et à mesure que l'on avance dans le test et que l'indice de difficulté des items augmente. De fait, il y a une relation très étroite entre les pondérations de chaque item sous le deuxième facteur, les indices de difficulté des items et les taux d'abstention. Plus l'indice de difficulté des items est élevé, moins les pondérations sous le deuxième facteur sont importantes et moins il y a d'abstention. La corrélation entre les pondérations sous le deuxième facteur et les indices de difficulté des items est de -0,918 ; la corrélation entre ces pondérations et les taux d'abstention à chaque item est de 0,823 ; la corrélation entre les indices de difficulté et les taux d'abstention, de -0,789 (voir le tableau 4, à l'annexe A).

Roznowski, Tucker et Humphreys (1991) expliquent qu'un test caractérisé par un facteur unique auquel s'ajoute de l'erreur aléatoire présentera un deuxième facteur dont le patron se distingue du premier, tant au niveau des pondérations qu'au niveau du signe qui les accompagne. Ils ajoutent que le deuxième facteur d'une matrice de corrélation basée sur une échelle de Guttman parfaite produit une courbe qui s'apparente à une ogive. Les items faciles et les items difficiles ont des pondérations élevées, mais de signe contraire. Les items de difficulté moyenne ont des pondérations presque nulles. C'est précisément ce qu'on observe avec le test analysé. Outre le fait que les items faciles et les items difficiles ont les pondérations les plus grandes et que celles-

ci sont de signe contraire, on note que les pondérations les plus grandes sous le deuxième facteur correspondent aux pondérations les plus faibles sous le premier facteur. De fait, l'indice de dimensionnalité obtenu à partir des pondérations du deuxième facteur est de 0,036. Cet indice est d'autant moins grand que le test est unidimensionnel.

Enfin, nous avons appliqué le test T de Stout (1987) pour vérifier deux hypothèses : celle d'un deuxième facteur totalement distinct du premier et celle d'un facteur temps suffisamment important pour rejeter l'hypothèse de l'unidimensionnalité. La première hypothèse a été vérifiée en appliquant le test T aux items qui ont les pondérations les plus grandes sous le deuxième facteur et les pondérations les plus faibles sous le premier facteur. La deuxième hypothèse a été vérifiée en appliquant le test T aux items dont le taux d'abstention est supérieur à 0,05, en arrondissant au centième le plus proche, pourvu que tous les items qui suivent aient également un taux d'abstention plus grand que 0,05. En raison du grand nombre de sujets dans l'échantillon, nous avons fixé le seuil de signification à 0,01. Les résultats montrent que le test de mathématiques peut admettre l'hypothèse de l'unidimensionnalité dans les deux cas (voir le tableau 5, à l'annexe A). Les résultats sont moins convaincants pour la deuxième hypothèse. Toutefois, le test T de Stout (1987) a été conçu pour s'appliquer à des items dont les pondérations sont élevées sous le deuxième facteur et faibles sous le premier facteur, ceci afin de voir ce qui arrive dans le cas d'items associés à des facteurs nettement distincts. Or, en considérant les items qui se suivent et dont le taux d'abstention est plus grand que 0,05, on a pris en compte des items dont les pondérations sont également assez élevées sous le premier facteur.

Nous avons également réalisé des analyses en composantes principales pour les hommes et pour les femmes pris séparément. Le but était de voir si le test présentait des caractéristiques

similaires pour chacun des groupes pris séparément. Que ce soit au niveau de la figure qui reproduit la valeur des racines latentes en fonction des facteurs, au niveau de l'écart entre la première racine latente et la deuxième, au niveau du rapport L ou au niveau de la proportion de variance expliquée par les premiers facteurs, on observe des tendances semblables pour chacun des groupes. Il y a un premier facteur prédominant et capable d'expliquer une proportion importante de la variance. Le deuxième facteur pourrait s'expliquer par de l'erreur liée à la difficulté des items et au taux d'abstention et résulter de la limite de temps imposée. Enfin le test T de Stout (1987) utilisé pour vérifier la présence d'un deuxième facteur distinct du premier montre que le test de mathématiques analysé admet l'hypothèse de l'unidimensionnalité. L'application du test T de Stout (1987) pour vérifier l'existence d'un facteur temps négligeable donne des résultats moins convaincants. Comme nous l'expliquions précédemment, le test T n'a pas été conçu à cette fin, et les items considérés ne sont pas tous nettement distincts de ceux qui contribuent le plus au premier facteur. Ceci dit, il n'en demeure pas moins que les résultats du test T ne sont pas significatifs au seuil de 0,01. Des différences existent entre les hommes et les femmes. Toutefois, celles-ci sont peu accentuées. De fait, la corrélation entre les pondérations de chaque groupe sous le premier facteur est de 0,946 et la corrélation entre les pondérations de chaque groupe sous le deuxième facteur est de 0,974. Étant donné que la cohérence interne du test est très semblable pour chacun des groupes pris séparément, que les tendances observées au niveau de la structure factorielle convergent, quel que soit le critère et quel que soit le groupe, et que le deuxième facteur est également lié à la difficulté statistique des items et au taux d'abstention, il nous apparaît que le test est unidimensionnel et, somme toute, également valide pour chacun des groupes comparés (voir les figures 2, 3, et 4 et les tableaux 1 à 5, à l'annexe A).

Analyse des données

L'étude de l'efficacité des méthodes d'analyse du fonctionnement différentiel des items exige que l'on sache quels items ont un fonctionnement différentiel et l'ampleur des différences dans le fonctionnement des items. Lorsqu'on utilise des données réelles et des échantillons de taille réduite, l'on ne connaît a priori ni l'un ni l'autre, d'où la nécessité de déterminer quels items ont un fonctionnement différentiel. Pour ce faire, nous étudierons le fonctionnement différentiel des items dans la population. La méthodologie adoptée pour répondre aux questions de recherche comporte donc trois volets : l'étude du fonctionnement différentiel des items dans la population, l'étude du fonctionnement différentiel des items dans des échantillons de taille réduite et l'étude de la fidélité et de l'efficacité des méthodes d'analyse du fonctionnement différentiel. À ces considérations s'ajoutent des précisions sur les programmes d'ordinateur utilisés pour effectuer les analyses de fonctionnement différentiel et les conditions dans lesquelles les analyses seront appliquées.

L'étude du fonctionnement différentiel des items dans la population

L'étude du fonctionnement différentiel des items dans la population vise à identifier les items qui ont un fonctionnement différentiel d'une ampleur prédéterminée. Parmi les méthodes empiriques et non paramétriques d'analyse du fonctionnement différentiel des items, trois fournissent une mesure de l'ampleur des différences dans le fonctionnement des items : la méthode de Mantel-Haenszel avec son rapport alpha et sa transformation en delta ou en Z de Mantel-Haenszel, la méthode de standardisation avec la différence de difficulté standardisée et la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a,

1992 b, 1993 et 1995). Nous utiliserons les trois méthodes. L'identification des items qui ont un fonctionnement différentiel non négligeable reposera sur un double critère : l'ampleur des différences de difficulté standardisée et la convergence des méthodes. La prise en considération de la convergence des méthodes a pour but de s'assurer que les items identifiés comme ayant un fonctionnement différentiel ne sont pas imputables à une particularité de la méthode.

Pour étudier le fonctionnement différentiel des items dans la population, nous nous servirons de l'échantillon de 40 000 sujets sélectionné pour déterminer les caractéristiques du test et en évaluer l'unidimensionnalité. L'utilisation de toute la population ciblée aurait pris trop d'espace sur l'ordinateur, et il n'était pas certain que les programmes utilisés pour étudier le fonctionnement différentiel puissent fonctionner. Outre ces considérations d'ordre pratique, trois raisons justifient le recours à un échantillon au lieu d'utiliser toute la population. L'échantillon est représentatif de la population, puisqu'il a été sélectionné de façon aléatoire et qu'il compte 25 pour cent de cette population. Les méthodes utilisées pour identifier les items qui accusent des différences de fonctionnement non négligeables reposent en dernière analyse sur des différences de proportions. Avec l'échantillon de 40 000 sujets, l'erreur-type d'une différence de proportions entre les deux groupes comparés ne peut dépasser 0,04 à chaque niveau d'habileté, en postulant une répartition uniforme des sujets à tous les niveaux d'habileté. Il en va de même pour tous les sujets dont le score se situe entre 11 et 43, là où se situent la majorité des sujets. Par ailleurs, l'erreur-type maximale théoriquement possible pour les différences de difficulté standardisée ne devrait pas dépasser les 0,005, en postulant, ici encore, une répartition uniforme des sujets à tous les niveaux d'habileté et en appliquant la formule de Dorans et Holland (1993) pour le groupe focalisé.

Seront considérés comme ayant un fonctionnement différentiel non négligeable, les items dont la différence de difficulté standardisée est égale ou supérieure à 0,05 en valeur absolue, en arrondissant au centième le plus proche. De plus, ces items devront avoir un indice qui les situe parmi les indices dont les différences de fonctionnement sont les plus grandes, en valeur absolue, lorsqu'on applique la méthode du delta de Mantel-Haenszel ou la méthode de standardisation modifiée. La grandeur des différences de difficulté standardisée retenue s'appuie sur les recommandations de Schmitt et Dorans (1990). Ces derniers affirment qu'un seuil de 0,05 permet d'identifier comme suspects un grand nombre d'items, la plupart s'avérant difficiles à expliquer à la suite d'un examen du contenu, tandis qu'un seuil de 0,10 permet d'identifier un petit nombre d'items dont les différences de fonctionnement peuvent facilement s'expliquer à la lecture des items. Pour cette raison, ils recommandent un seuil de 0,05 pour la recherche et un seuil de 0,10 dans la pratique. Nous retenons le seuil de 0,05, parce que le test analysé a déjà fait l'objet d'un contrôle pour minimiser et équilibrer les différences de fonctionnement des items. Nous présumons que les différences de fonctionnement des items seront plutôt petites, hypothèse soutenue par l'étude de Pang, Tian et Boss (1994). Nous nous assurerons que la grandeur des différences de fonctionnement retenue dépasse les variations possibles à l'intérieur d'un groupe en appliquant la méthode des groupes aléatoires proposée par Shepard, Camilli et Williams (1984) et, plus récemment, par Hambleton et Rogers (1991).

L'étude du fonctionnement différentiel des items dans les échantillons

L'étude du fonctionnement différentiel des items dans les échantillons de taille réduite porte exclusivement sur la méthode de Mantel-Haenszel, la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995) et la

méthode de régression logistique. Les indices retenus pour identifier les items qui ont un fonctionnement différentiel et étudier la fidélité et l'efficacité relative des méthodes sont le delta et le khi carré de Mantel-Haenszel, la différence de difficulté standardisée modifiée et le khi carré d'amélioration pour la régression logistique.

Pour répondre aux questions de recherche, nous appliquerons les méthodes d'analyse du fonctionnement différentiel des items à des échantillons de 250, de 500, de 1000 et de 2000 sujets. Les échantillons seront choisis parmi les 40 000 sujets retenus pour étudier le fonctionnement différentiel des items dans la population. Pour chaque grandeur d'échantillon, nous reprendrons l'échantillonnage 100 fois. Dans chaque cas, nous prélèverons un nombre égal de personnes de chaque sexe et la sélection des sujets sera effectuée de façon aléatoire et indépendante. Les échantillons de 250 sujets compteront donc 125 hommes et 125 femmes ; les échantillons de 500 sujets, 250 hommes et 250 femmes ; ceux de 1000 sujets, 500 femmes et 500 hommes ; et ceux de 2000 sujets, 1000 hommes et 1000 femmes. Deux remarques s'imposent. L'une concerne l'utilisation de groupes égaux. L'autre a trait à la sélection des sujets pour les analyses répétées.

L'utilisation de groupes égaux au lieu de groupes proportionnels tient à plusieurs facteurs. Tout d'abord, il faut noter que l'utilisation de groupes égaux ou disproportionnés a peu d'effet sur la mesure des différences de fonctionnement des items à chaque niveau d'habileté et sur la détection des items qui ont un fonctionnement différentiel, lorsqu'il y a peu de différence dans la proportion des groupes comparés, c'est-à-dire lorsque les proportions se situent entre 40 et 60 pour cent. Ceci dit, eu égard au but des analyses de fonctionnement différentiel, il nous apparaît préférable d'accorder une égale importance à chacun des groupes comparés. Lorsqu'on prélève

des échantillons de taille réduite dans deux sous-populations d'habileté inégale, comme c'est le cas ici, l'utilisation de groupes égaux offre une meilleure assurance d'avoir un nombre suffisant de sujets de chaque groupe à tous les niveaux d'habileté. Qui plus est, le procédé situe la recherche dans la lignée des autres recherches. Enfin, Cohen (1988) détermine la puissance d'un test statistique en fonction de la grandeur de l'effet et du nombre de sujets à partir de groupes égaux et il utilise la moyenne harmonique pour déterminer celle-ci lorsque les groupes sont disproportionnés. Dans l'état actuel de la recherche, nous ne savons pas dans quelle mesure le procédé s'applique à des analyses de fonctionnement différentiel, mais il nous paraît plus sage d'utiliser des groupes égaux dans un premier temps.

La répétition des analyses de fonctionnement différentiel des items répond à un double objectif : déterminer la fidélité des méthodes et évaluer leur efficacité tout en tenant compte des fluctuations possibles d'échantillonnage. L'expression désigne ici les fluctuations issues d'une combinaison différente de sujets, que ceux-ci soient répartis différemment d'un échantillon à l'autre ou qu'ils soient totalement distincts. Pour tenir compte des fluctuations possibles d'échantillonnage, nous avons prévu sélectionner les sujets comme si nous reprenions l'étude à zéro à chaque fois. Ce faisant, les sujets ont toujours la même probabilité de sélection pour une même grandeur d'échantillon. Toutefois, parce que nous disposons d'un nombre limité de sujets pour effectuer la sélection, les mêmes sujets pourront revenir dans plusieurs échantillons, mais combinés différemment. Il y aura donc une certaine proportion de sujets communs à deux échantillons.

Nous avons prévu sélectionner les échantillons de taille réduite parmi les 40 000 sujets représentant la population d'origine. Dans ces conditions, la sélection d'un échantillon de 2000

sujets représente cinq pour cent des sujets disponibles, six pour cent parmi les hommes et quatre pour cent parmi les femmes, puisque nous utiliserons des groupes égaux. En appliquant la méthode décrite par Moroney (1970) pour estimer le nombre de combinaisons possibles, il appert que ce nombre est astronomique pour ne pas dire infini. Il en résulte que la probabilité de sélectionner deux échantillons identiques est, à toutes fins utiles, inexistante. Par contre, on peut s'attendre à ce que le même sujet réapparaisse dans quatre ou six échantillons différents. Ces derniers seront toutefois combinés différemment à d'autres sujets. Dans cette optique, on peut prévoir cinq pour cent de sujets communs à deux échantillons consécutifs. La sélection d'échantillons plus petits donne lieu à moins de combinaisons. Toutefois, le nombre demeure considérable, et la proportion de sujets communs à deux échantillons diminue. Le tableau 6, à l'annexe A, fait état de la probabilité de sélection d'un sujet pour chaque grandeur d'échantillon et de la proportion de sujets communs à deux échantillons. En gros, on peut dire que la procédure d'échantillonnage s'avère plus problématique pour les échantillons de 2000 sujets qu'elle ne l'est pour les échantillons de 1000, de 500 ou de 250 sujets, mais elle demeure dans les limites de l'acceptable. Par ailleurs, la présente étude porte sur des méthodes d'analyse du fonctionnement différentiel applicables à des échantillons de taille réduite. Dans cette optique, la sélection d'un échantillon de 2000 sujets constitue le point ultime pour étudier les tendances.

L'identification des items qui ont un fonctionnement différentiel dans les échantillons de taille réduite sera basée sur la valeur des indices à un centile prédéterminé ou sur une valeur critique fixée a priori. Dans le cas d'un centile prédéterminé, il s'agira des centiles C_{95}, C_{90} ou C_{80} pour le khi carré de Mantel-Haenszel et le khi carré d'amélioration et des centiles $C_{2,5}$ et $C_{97,5}$, C_5 et C_{95} ou C_{10} et C_{90} pour le delta de Mantel-Haenszel et la différence de difficulté standardisée modifiée. Dans le cas d'une valeur critique fixée a priori, il s'agira d'une valeur usuelle : la

valeur critique du khi carré de Mantel-Haenszel au seuil de 0,05, à un degré de liberté ; la valeur critique du khi carré d'amélioration au seuil de 0,05, à deux degrés de liberté pour la régression logistique ; et l'unité pour le delta de Mantel-Haenszel. Dans le cas de la différence de difficulté standardisée modifiée, il n'y a pas de valeur recommandée fixée a priori. Nous retiendrons la valeur d'indice égale ou supérieure à 1,96 écart-type et la valeur d'indice égale ou inférieure à -1,96 écart-type pour l'ensemble des distributions. Pour déterminer la valeur d'indice, nous nous référerons à chaque distribution d'indices moyens obtenue pour les 100 échantillons de 250, de 500, de 1000 et de 2000 sujets ainsi qu'à la distribution des indices pour la population.

Les programmes utilisés pour étudier le fonctionnement différentiel des items

Nous utilisons les mêmes programmes pour étudier le fonctionnement différentiel des items dans les échantillons de taille réduite et le fonctionnement différentiel des items dans la population. Pour l'étude du fonctionnement différentiel des items à l'aide de la méthode de Mantel-Haenszel, nous nous servons d'un programme préparé par Rogers et Hambleton (1994) pour micro-ordinateur. L'étude du fonctionnement différentiel des items à l'aide de la méthode de standardisation est effectuée à l'aide du même programme, ce dernier fournissant simultanément les valeurs calculées du khi carré de Mantel-Haenszel, du delta de Mantel-Haenszel et de la différence de difficulté standardisée. Pour l'étude du fonctionnement différentiel des items à l'aide de la méthode de standardisation modifiée, nous nous servons de la version 1996 de Testgraf, un programme mis au point par Ramsay (1995). Quant à l'étude du fonctionnement différentiel des items à l'aide de la méthode de régression logistique, nous nous servons de la version 1994 du programme SPSS pour micro-ordinateur et de la procédure LOGISTIC REGRESSION (Norušis, 1990).

Des précisions s'imposent en ce qui concerne l'étude du fonctionnement différentiel des items avec la méthode de régression logistique et la procédure LOGISTIC REGRESSION. Deux approches peuvent être utilisées pour identifier les items qui ont un fonctionnement différentiel : postuler que la probabilité de répondre correctement à un item dépend de l'habileté des sujets et de leur appartenance à un groupe ou postuler qu'elle dépend à la fois de leur habileté, de leur appartenance à un groupe et de l'interaction entre l'habileté des sujets et leur appartenance à un groupe, puis vérifier si les coefficients de régression logistique associés à une différence dans le fonctionnement des items sont significativement différents de zéro. Ne prenant pas en compte l'effet d'interaction, la première approche s'apparente davantage à la méthode de Mantel-Haenszel, à la méthode de standardisation et à la méthode de standardisation modifiée. Moins puissante que la première, la deuxième approche est la seule à prendre en compte l'effet d'interaction et à permettre, dans un second temps, de déterminer le type de fonctionnement différentiel qui caractérise un item. Nous utilisons la deuxième approche, d'une part parce qu'elle correspond à l'approche préconisée par Swaminathan et Rogers (1990), d'autre part, parce qu'elle permet la comparaison avec d'autres recherches qui utilisent cette approche.

Une autre remarque s'impose. Swaminathan et Rogers (1990) ont mis au point un programme qui vérifie si, pris globalement, les coefficients de régression logistique associés au groupe et à l'interaction entre le groupe et l'habileté des sujets sont significatifs. SPSS ne dispose pas de ce programme. Pour vérifier si les coefficients de régression logistique sont significatifs ou non, nous procédons comme suit. Nous postulons un modèle sans fonctionnement différentiel :

$$P(u=1) = \frac{e^{z}}{\left[1 + e^{z}\right]} \qquad (22)$$

où

$$z = \beta_0 + \beta_1\theta \ . \tag{23}$$

β_0 est une constante, et $\beta_1\theta$, le coefficient de régression logistique associé à l'habileté des sujets.

Nous postulons également un modèle avec fonctionnement différentiel uniforme ou non uniforme :

$$P(u=1) = \frac{e^z}{\left[1 + e^z\right]} \tag{24}$$

où

$$z = \beta_0 + \beta_1\theta + \beta_2 G + \beta_3(\theta G) \ . \tag{25}$$

β_0 est une constante ; $\beta_1\theta$, le coefficient de régression logistique associé à l'habileté des sujets, $\beta_2 G$, le coefficient de régression associé à l'appartenance à un groupe ; et $\beta_3(\theta G)$, le coefficient de régression associé à l'interaction entre l'habileté des sujets et leur appartenance à un groupe.

Nous déterminons ensuite le khi carré associé à chacun des modèles et nous calculons la différence entre les khis carrés de chaque modèle. Cette différence correspond à ce qu'il est convenu d'appeler le khi carré d'amélioration. Tandis que les khis carrés de chaque modèle indiquent si tous les coefficients de régression du modèle postulé sont significatifs à un seuil prédéterminé, le khi carré d'amélioration indique si les coefficients de régression ajoutés au modèle sans fonctionnement différentiel sont significatifs. La distribution du khi carré d'amélioration correspond à celle du khi carré lorsque l'hypothèse nulle est vraie. Pour la

présente étude, un item est dit avoir un fonctionnement différentiel si le khi carré d'amélioration est significatif à deux degrés de liberté, le modèle avec fonctionnement différentiel comptant deux variables de plus que le modèle sans fonctionnement différentiel.

Les conditions d'analyse du fonctionnement différentiel des items

D'autres précisions s'imposent en ce qui concerne les analyses de fonctionnement différentiel effectuées. Tout d'abord, il convient de signaler que nous utilisons un critère interne comme mesure de l'habileté des sujets en mathématiques : les résultats au test analysé. Deux raisons motivent ce choix. Les résultats au test sont la mesure la plus juste et la plus fiable dont on dispose de l'habileté des sujets en mathématiques. Lors de l'élaboration d'un test, les résultats au test sont généralement la seule mesure disponible de l'habileté que l'on prétend mesurer.

Nous utilisons comme mesure de l'habileté des sujets la somme des réponses correctes pour l'ensemble des items qui composent le test ou, dans le cas de la méthode de standardisation modifiée par l'application du modèle de Ramsay, le score latent estimé pour la totalité des items. L'utilisation du score total au test pour estimer l'habileté des sujets est justifiée par le fait que le test admet l'hypothèse de l'unidimensionnalité, que ce score est plus fiable qu'un score qui utiliserait moins d'items et qu'il s'avère somme toute également valide pour chacun des groupes comparés.

Nous n'éliminerons pas l'item étudié de l'estimation d'habileté des sujets. Nous suivons en cela la recommandation de Donohue, Holland et Thayer (1993). Ces derniers soutiennent que l'inclusion de l'item étudié a pour effet de réduire le nombre de faux positifs lorsqu'on utilise la

méthode de Mantel-Haenszel, d'où leur recommandation, même si l'erreur-type des deltas de Mantel-Haenszel est moins précise. Mis à part cette recommandation, deux autres raisons motivent notre décision. L'élimination de l'item étudié du score total s'avère très complexe dans le cas de la méthode de régression logistique, et nous voulons maintenir les mêmes conditions pour toutes les méthodes. Enfin, des analyses préliminaires donnent à penser qu'il y a à peu près autant d'items qui favorisent les femmes qu'il y en a qui favorisent les hommes. Dans de telles conditions, les différences qui peuvent exister dans le fonctionnement des items sont neutralisées au niveau du score total.

En ce qui concerne la méthode de Mantel-Haenszel, nous effectuerons les analyses en utilisant autant de strates qu'il y a de scores possibles, suivant en cela une recommandation faite par Clauser, Mazor et Hambleton (1991 a). En ce qui concerne la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995), nous considérerons les réponses omises comme une sixième réponse possible. Par ailleurs, nous fixerons à 61 le nombre de points d'estimation de la probabilité des réponses en fonction de l'habileté des sujets et nous maintiendrons ce nombre pour toutes les grandeurs d'échantillon. Nous laisserons au programme, Testgraf, le soin de fixer la grandeur de l'intervalle pour l'application du procédé de lissage. Le programme fixe celle-ci de manière à optimiser le procédé. Pour ce faire, il prend en compte la grandeur de l'échantillon. L'intervalle variera donc en fonction de la taille des échantillons, mais il sera constant pour tous les échantillons de même taille.

L'étude de la fidélité et de l'efficacité des méthodes

La fidélité des méthodes sera évaluée en se basant sur la stabilité des indices d'un échantillon à un autre de même taille. Celle-ci sera donnée par la corrélation entre les indices de même nature pour les échantillons de même taille pris deux à deux. S'y ajoutera le nombre de fois qu'un item est identifié comme ayant un fonctionnement différentiel dans les échantillons de même taille. L'efficacité relative des méthodes sera évaluée par la comparaison des taux de détection des items qui ont un fonctionnement différentiel non négligeable dans la population et la comparaison des taux de détection des items identifiés à tort comme ayant un fonctionnement différentiel. À titre complémentaire, nous y ajouterons la validité des indices. Le terme est ici réservé à la relation entre les indices de fonctionnement différentiel des items obtenus dans les échantillons et les items identifiés comme ayant un fonctionnement différentiel non négligeable dans la population. Pour ce faire, nous examinerons le nombre de décisions correctes dans chaque échantillon et la relation entre les items détectés dans les échantillons et les items retenus comme ayant un fonctionnement non négligeable dans la population.

CHAPITRE V

RÉSULTATS

Les pages qui suivent font état des résultats des analyses de fonctionnement différentiel dans la population, des résultats des analyses de fonctionnement différentiel dans les échantillons et des résultats des études sur la fidélité et l'efficacité des méthodes d'analyse du fonctionnement différentiel des items. Nous examinerons d'abord le fonctionnement différentiel des items dans la population et, ensuite, le fonctionnement différentiel des items dans les échantillons.

Étude du fonctionnement différentiel des items dans la population

À titre de rappel, le fonctionnement différentiel des items dans la population est évalué à partir d'un ensemble de 40 000 sujets représentant la population ciblée. On y compte 16 521 personnes de sexe masculin et 23 479 personnes de sexe féminin. Les premières forment le groupe de référence ; les secondes, le groupe focalisé. Les méthodes et les indices utilisés sont la méthode de standardisation avec la différence de difficulté standardisée (DDS), la méthode de Mantel-Haenszel avec le delta de Mantel-Haenszel (Δ_{MH}) et la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993, 1995) avec la différence de difficulté standardisée modifiée (DDSM). Les différences de difficulté standardisée (DDS) servent à identifier les items qui ont un fonctionnement différentiel et fournissent une mesure de l'ampleur des différences dans le fonctionnement des items. Les deltas de Mantel-Haenszel (Δ_{MH}) et les différences de difficulté standardisée modifiée (DDSM) sont utilisés pour s'assurer que les items identifiés comme ayant un fonctionnement différentiel ne sont pas la résultante d'une particularité méthodologique.

Le tableau 2 résume les résultats des analyses de fonctionnement différentiel effectuées. Nous y indiquons la moyenne, la variance et l'écart-type ainsi que la valeur minimale, la valeur maximale et la médiane des indices obtenus pour chaque méthode. Pour fin de comparaison, le tableau 3 reproduit les mêmes données transformées en scores standardisés. Moyenne, variance et écart-type ne sont pas indiqués, puisque la valeur est identique pour les trois méthodes : 0 pour la moyenne et 1 pour la variance et l'écart-type. La figure 1 dépeint la répartition des indices pour chacune des trois méthodes.

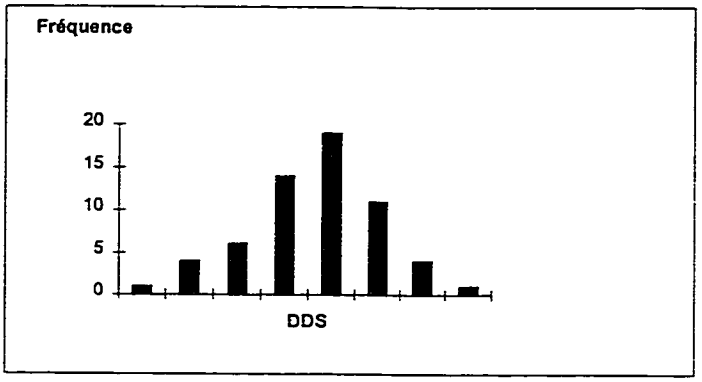
Tableau 2. Description des résultats des analyses de fonctionnement différentiel effectuées pour les 40 000 sujets représentant la population (N = 60 items).

Indices	**DDS**	Δ_{MH}	**DDSM**
Moyenne	0,00000	0,00933	-0,00003
Variance	0,00154	0,27098	0,00138
Écart-type	0,03922	0,52056	0,03712
Minimum	-0,091	-1,200	-0,080
Maximum	0,108	1,410	0,101
Médiane	0,004	0,055	0,0005

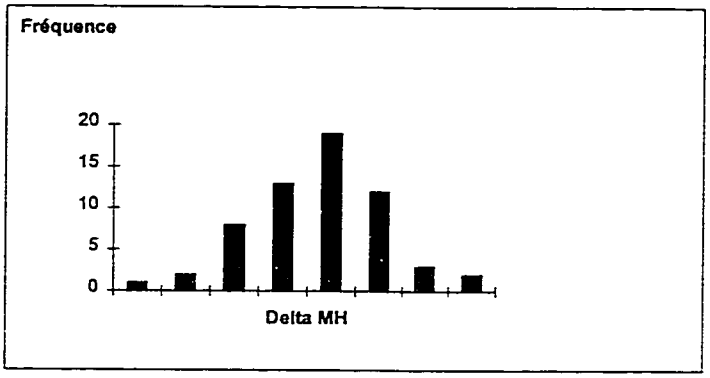
Tableau 3. Description des résultats en scores standardisés des analyses de fonctionnement différentiel effectuées pour les 40 000 sujets représentant la population (N = 60 items).

Indices	**DDS**	Δ_{MH}	**DDSM**
Minimum	-2,32025	-2,32314	-2,15438
Maximum	2,75370	2,69069	2,72193
Médiane	0,10199	0,08773	0,01437

DDS	Fréquence
-0,091	1
-0,063	4
-0,034	6
-0,006	14
0,023	19
0,051	11
0,08	4
plus grand	1



Delta MH	Fréquence
-1,2	1
-0,827	2
-0,454	8
-0,081	13
0,291	19
0,664	12
1,037	3
plus grand	2



DDSM	Fréquence
-0,08	1
-0,054	5
-0,028	6
-0,002	13
0,023	22
0,049	8
0,075	4
plus grand	1

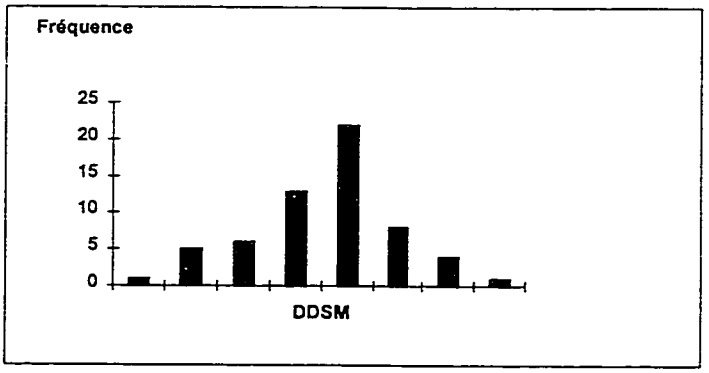


Figure 1. Distribution de fréquence des indices de fonctionnement différentiel pour la méthode de standardisation, la méthode du delta de Mantel-Haenszel et la méthode de standardisation modifiée.

Outre la très grande ressemblance qui existe entre les trois distributions d'indices, trois remarques s'imposent. Les différences de difficulté standardisée (DDS) ont une moyenne nulle. La moyenne des deltas de Mantel-Haenszel (Δ_{MH}) et celle des différences de difficulté standardisée modifiée (DDSM) sont presque nulles. La valeur minimale et la valeur maximale des indices suggèrent une répartition à peu près égale des indices en faveur de chaque groupe. La figure 1 le confirme. À partir de ces données, on comprend aisément que les items en faveur d'un groupe sont neutralisés par les items en faveur de l'autre groupe. Cette particularité montre le bien-fondé de l'utilisation du score total au test comme mesure de l'habileté des sujets.

Le tableau 4 reproduit les corrélations observées entre les différences de difficulté standardisée (DDS), les deltas de Mantel-Haenszel (Δ_{MH}) et les différences de difficulté standardisée modifiée (DDSM). Les corrélations sont extrêmement élevées et hautement significatives. Ces données attestent d'une très grande convergence entre les méthodes, une convergence qui n'étonne pas, puisque les trois méthodes poursuivent un même but et qu'elles reposent sur des assises communes.

Tableau 4. Corrélations de Pearson entre les indices de fonctionnement différentiel pour les 40 000 sujets représentant la population cible (N = 60 items).

Indices	**DDS**	Δ_{MH}	**DDSM**
DDS	1,00000*		
Δ_{MH}	0,98614*	1,00000*	
DDSM	0,99158*	0,97173*	1,00000*

* Le coefficient de corrélation est significatif au seuil de 0,01.

Nous avons convenu d'identifier comme ayant un fonctionnement différentiel tout item dont les différences de difficulté standardisée (DDS) étaient égales ou supérieures à 0,05, en valeur absolue, en arrondissant au centième le plus proche. Quinze items répondent à ce critère. Ce sont les items 4, 7, 9, 14, 17, 18, 19, 23, 29, 32, 34, 37, 38, 52 et 58. Parmi ces items, 13 accusent les différences de fonctionnement les plus grandes en valeur absolue, lorsqu'on se réfère aux deltas de Mantel-Haenszel (Δ_{MH}) et aux différences de difficulté standardisée modifiée (DDSM). Il s'agit des items 4, 7, 9, 14, 17, 19, 23, 29, 32, 37, 38, 52 et 58. L'item 34 se classe parmi les 15 items qui affichent les différences de fonctionnement les plus grandes en valeur absolue lorsqu'on se réfère aux deltas de Mantel-Haenszel (Δ_{MH}), mais non lorsqu'on se reporte aux différences de difficulté standardisée modifiée (DDSM). L'item 18 se classe parmi les 15 items qui accusent les différences de fonctionnement les plus grandes en valeur absolue lorsqu'on se réfère aux différences de difficulté standardisée modifiée (DDSM), mais non lorsqu'on consulte les deltas de Mantel-Haenszel (Δ_{MH}). Parce que tous ces items ont une différence de difficulté standardisée de 0,05 en valeur absolue, après arrondissement au centième le plus proche, et qu'il y a convergence avec au moins une autre méthode, ils seront désormais considérés comme les items avec un fonctionnement différentiel non négligeable dans la population.

Le tableau 5 fait état du rang des items avec un fonctionnement différentiel non négligeable et de la grandeur des indices, en valeur absolue, pour chacune des trois méthodes considérées. Le tableau 6 reproduit les paramètres d'items de ces 15 items et le tableau 7 fait état de la moyenne, de l'écart-type et de la médiane des différences de difficulté standardisée (DDS) pour les 15 items pris globalement et pour les items répartis en trois catégories selon la grandeur des différences de fonctionnement observées. Moyenne, médiane et écart-type ont été calculés à

partir des valeurs absolues. Fait à noter, il y a plus d'accord entre les méthodes pour les items qui affichent les différences de fonctionnement les plus grandes qu'il y en a pour les items qui affichent les différences de fonctionnement les moins grandes. Par ailleurs, à l'examen des figures fournies par Testgraf, il appert qu'un très grand nombre d'items ont des courbes de réponses correctes qui se croisent une ou plusieurs fois, généralement aux extrémités de l'échelle d'habileté.

Tableau 5. Rang et valeur des indices de fonctionnement différentiel, en valeur absolue, des 15 items avec un fonctionnement différentiel non négligeable pour chacune des trois méthodes d'analyse du fonctionnement différentiel utilisées (N = 40 000 sujets).

Item	Valeur absolue des indices			Rang des items basé sur la valeur absolue des indices *		
	DDS	Δ_{MH}	DDSM	DDS	Δ_{MH}	DDSM
4	0,048	0,73	0,050	13	9	11
7	0,053	0,69	0,060	11	12	8
9	0,058	0,71	0,059	10	10	9
14	0,077	1,12	0,074	3	3	4
17	0,091	1,01	0,080	2	4	2
18	0,046	0,52	0,047	15	19,5	13,5
19	0,108	1,41	0,101	1	1	1
23	0,071	0,75	0,076	4	8	3
29	0,059	0,88	0,054	9	6	10
32	0,061	0,70	0,049	8	11	12
34	0,047	0,64	0,038	14	13,5	19,5
37	0,052	0,61	0,045	12	15	15
38	0,067	0,91	0,062	7	5	6
52	0,068	0,80	0,066	6	7	5
58	0,070	1,20	0,061	5	2	7

* Le rang indiqué est celui de l'item lorsqu'on considère la totalité des items qui composent le test. Pour les Δ_{MH}, l'item 57 occupe le 13,5e rang avec l'item 34. Pour les DDSM, c'est l'item 10 qui occupe le 13,5e rang avec l'item 18. L'item 57 occupe le 29e rang lorsqu'on utilise les DDS ; l'item 10, le 17,5e rang. L'item 10 partage le 17,5e rang avec l'item 48.

Tableau 6. Paramètres des 15 items qui ont un fonctionnement différentiel non négligeable dans la population (N = 40 000 sujets).

Item	Paramètres d'item*			Corrélation bisériale*	Difficulté
	a	b	c		
4	1,091	-0,566	0,248	0,464	0,734
7	0,876	-0,511	0,343	0,397	0,691
9	0,964	-0,030	0,190	0,484	0,594
14	1,268	-0,019	0,077	0,596	0,540
17	0,633	0,134	0,244	0,369	0,528
18	0,702	0,519	0,114	0,439	0,468
19	0,914	0,391	0,061	0,542	0,458
23	0,725	0,074	0,274	0,367	0,550
29	1,192	0,074	0,049	0,600	0,488
32	0,613	0,641	0,171	0,323	0,405
34	1,126	0,066	0,122	0,547	0,470
37	0,721	0,798	0,165	0,366	0,410
38	1,081	0,368	0,124	0,532	0,378
52	0,598	1,296	0,174	0,258	0,301
58	0,715	1,339	0,086	0,334	0,200

* Les paramètres d'items ont été estimés à l'aide du modèle de Ramsay et du programme Testgraf.

Tableau 7. Différences de difficulté standardisée (DDS), en valeur absolue, pour les items qui ont un fonctionnement différentiel non négligeable dans la population (N = 40 000 sujets).

Groupe d'items	Liste des items	Nombre d'items	Moyenne	Écart-type	Médiane
Totalité des items (0,046 - 0,108)	4, 7, 9, 14, 17, 18, 19 23, 29, 32, 34, 37, 38, 52, 58.	15	0,06507	0,01718	0,06100
Items avec un FD assez élevé (0,070 - 0,108)	14, 17, 19, 23, 58.	5	0,08340	0,01610	0,07700
Items avec un FD faible (0,058 - 0,068)	9, 29, 32, 38, 52.	5	0,06260	0,00462	0,06100
Items avec un FD très faible (0,046 - 0,053)	4, 7, 18, 34, 37.	5	0,04920	0,00311	0,04800

Pour s'assurer que les 15 items identifiés comme ayant un fonctionnement différentiel non négligeable affichent des différences de fonctionnement plus grandes que celles auxquelles on peut s'attendre entre deux groupes d'une même sous-population, nous avons appliqué la méthode des groupes aléatoires. Pour ce faire, nous avons réparti les hommes en deux groupes égaux de façon aléatoire (H1 et H2). Nous avons fait de même avec les femmes (F1 et F2). Nous avons ensuite procédé à une analyse du fonctionnement différentiel des items avec les deux groupes de sexe masculin (H1 et H2). Nous avons répété l'opération avec les deux groupes de sexe féminin (F1 et F2). Seule la méthode de standardisation a été appliquée, puisque ce sont les différences de difficulté standardisée qui servent de critère. Le tableau 8 reproduit les résultats des analyses de fonctionnement différentiel effectuées.

Tableau 8. Description des résultats des analyses de fonctionnement différentiel effectuées entre les groupes de même sexe avec la méthode de standardisation et les différences de difficulté standardisée (N = 60 items).

Indices	**DDS entre les groupes de sexe masculin (H1 et H2)** (N = 16 520 sujets)	**DDS entre les groupes de sexe féminin (F1 et F2)** (N = 23 478 sujets)
Moyenne	0,00007	0,00003
Variance	0,00017	0,00016
Écart-type	0,01311	0,01278
Minimum	-0,027	-0,031
Maximum	0,032	0,028
Médiane	0,001	-0,002

Comme on peut le constater, l'étendue des différences de difficulté standardisée (DDS) entre les groupes de même sexe est nettement moins grande que celle qui a été observée entre les groupes de sexe différent, et ce, même si le nombre de sujets dans les groupes comparés est plus petit.

Chez les hommes, les différences de difficulté standardisée (DDS) varient de - 0,027 à 0,032 ; chez les femmes, de - 0,031 à 0,028. Aucune n'atteint le seuil de 0,05 en valeur absolue, ni même celui de 0,046, la valeur minimale requise avant arrondissement pour considérer qu'un item a un fonctionnement différentiel non négligeable. De fait, les 15 items considérés comme ayant un fonctionnement différentiel non négligeable se situent tous à plus de trois écarts-types au-dessus ou au-dessous de la moyenne des indices entre les groupes de même sexe. Nous en déduisons que les différences de fonctionnement des 15 items retenus sont valides et qu'elles reflètent des différences liées au genre des individus.

En appliquant le critère des trois écarts-types, nous aurions pu y inclure cinq items de plus : les items 6, 10, 15, 16 et 48. Ces items accusent des différences de difficulté standardisée (DDS) égales ou supérieures à 0,04 en valeur absolue, après arrondissement au centième le plus proche. De plus, ils se situent parmi les 21 items qui ont le fonctionnement différentiel le plus grand en valeur absolue lorsqu'on se réfère soit aux différences de difficulté standardisée modifiée (DDSM) pour l'item 15, soit aux deltas de Mantel-Haenszel (Δ_{MH}) et aux différences de difficulté standardisée modifiée (DDSM) pour les items 6, 10, 16 et 48. Un autre item, l'item 22, présente une différence de difficulté standardisée (DDS) de 0,04. Cet item ne se classe pas parmi les 21 items qui ont les indices les plus grands en valeur absolue lorsqu'on applique la méthode du delta de Mantel-Haenszel ou la méthode de standardisation modifiée. Pour cette raison, il ne pourrait être retenu.

Nous préférons nous en tenir aux recommandations de Schmitt et Dorans (1990) et ne retenir que les items qui accusent des différences de fonctionnement égales ou supérieures à 0,05 en valeur absolue, après arrondissement au centième le plus proche. D'une part, il n'y a pas d'études qui

puissent servir de guide. D'autre part, un écart-type sépare la valeur d'indice la plus grande des variations à l'intérieur des groupes (0,032 en valeur absolue) de la valeur la plus petite des items retenus comme ayant un fonctionnement différentiel non négligeable (0,046 en valeur absolue, avant arrondissement au centième le plus proche). Le seuil retenu se situe en fait à 3,50 écarts-types de la moyenne des indices issus des comparaisons entre les groupes du même sexe. Le risque de retenir des items dont les différences de fonctionnement ne sont pas liées au genre des individus est encore moins grand que si nous fixions le seuil à trois écarts-types. Nous sommes cependant conscients que cette décision peut donner lieu à des taux plus élevés de faux positifs. Ce sont là les limites d'une étude effectuée avec des données réelles, mais qui reflètent des conditions qui peuvent se réaliser dans la pratique.

Nous avons également appliqué la méthode des groupes aléatoires pour nous assurer que les items retenus comme ayant un fonctionnement différentiel non négligeable étaient fiables. Pour ce faire, nous avons réuni un des deux groupes masculins à un des deux groupes féminins constitués de façon aléatoire pour vérifier la validité de ces mêmes items (H1 et F1), puis nous avons réuni les deux groupes qui restaient (H2 et F2). Nous avons ensuite appliqué la méthode de standardisation à chacun de ces deux sous-ensembles. La méthode n'est pas idéale. Il y a moitié moins de sujets dans chacun des sous-ensembles qu'il y en a lorsqu'on considère la totalité des 40 000 sujets représentant la population. Les sous-ensembles sont donc susceptibles d'offrir des indices moins fiables. Toutefois, dans la mesure où l'on parvient à identifier les mêmes items que ceux qui ont été retenus à partir de la totalité des 40 000 sujets représentant la population, on pourra accorder davantage foi dans la justesse des items retenus comme ayant un

fonctionnement différentiel non négligeable. C'est l'interprétation des résultats divergents qui fait problème.

Le tableau 9 fait état des résultats des analyses de fonctionnement différentiel effectuées avec l'ensemble de 40 000 sujets scindé en deux de façon aléatoire. Ces derniers montrent que la variabilité des indices issus des comparaisons entre groupes de sexe différent est plus grande que la variabilité des indices issus des comparaisons entre groupes de même sexe.

Tableau 9. Description des résultats des analyses de fonctionnement différentiel effectuées entre des groupes de sexe différent avec la méthode de standardisation et les différences de difficulté standardisée (N = 60 items).

Indices	**DDS entre deux groupes de sexe différent (H1 et F1)** (N = 20 000 sujets)	**DDS entre deux groupes de sexe différent (H2 et F2)** (N = 20 000 sujets)
Moyenne	0,00000	-0,00238
Variance	0,00166	0,00189
Écart-type	0,04072	0,04353
Minimum	-0,103	-0,160
Maximum	0,114	0,103
Médiane	0,002	0,0025

La corrélation entre les différences de difficulté standardisée (DDS) des deux échantillons est de 0,888. En ne retenant comme ayant un fonctionnement différentiel non négligeable que les items qui accusent des différences de difficulté standardisée (DDS) de 0,05 en valeur absolue après arrondissement au centième le plus proche, il y a 16 items avec un fonctionnement différentiel dans le premier échantillon (H1 et F1). Ce sont les items 4, 7, 9, 14, 16, 17, 18, 19, 23, 29, 32, 34, 37, 38, 52 et 58. Seul l'item 16 ne figure pas parmi les 15 items initialement retenus. Cet

item se situe tout juste sous le seuil de décision lorsqu'on se base sur l'ensemble de 40 000 sujets. La deuxième comparaison donne des résultats qui diffèrent davantage. On y détecte 15 items avec un fonctionnement différentiel non négligeable : les items 4, 6, 7, 9, 14, 17, 19, 22, 23, 29, 32, 37, 38, 52 et 58. L'item 16 ne figure pas dans la liste cette fois-ci. Treize items apparaissent parmi les 15 items initialement retenus comme ayant un fonctionnement différentiel non négligeable. Il s'agit des items 4, 7, 9, 14, 17, 19, 23, 29, 32, 37, 38, 52 et 58. Deux items n'y apparaissent pas : les items 6 et 22. Par ailleurs, deux items qui s'avèrent avoir un fonctionnement différentiel lorsqu'on se base sur la totalité des 40 000 sujets n'ont pas été détectés : les items 18 et 34. Ces deux items accusaient moins de convergence que les autres. À partir des analyses effectuées, nous en concluons que 13 des 15 items retenus sont fiables et valides : les items 4, 7, 9, 14, 17, 19, 23, 29, 32, 37, 38, 52 et 58. Deux items sont plus problématiques : les items 18 et 34. Nous les retenons malgré tout. Les données issues de la totalité des 40 000 sujets sont théoriquement plus fiables que celles issues d'un sous-échantillon de 20 000 sujets. De plus, ils satisfont au critère que nous nous étions fixé dans un des deux sous-échantillons. Par contre, nous voyons là une raison de plus de ne pas retenir plus de 15 items comme ayant un fonctionnement différentiel non négligeable. En retenir davantage ne ferait qu'accroître le nombre et la proportion d'items problématiques.

Étude du fonctionnement différentiel des items dans les échantillons

Ainsi que nous l'indiquions dans les chapitres précédents, la présente étude a pour but de comparer la fidélité et l'efficacité de méthodes d'analyse du fonctionnement différentiel des items applicables à des échantillons de taille réduite. À cette fin, nous avons prélevé 400 échantillons de sujets : 100 échantillons de 250, de 500, de 1000 et de 2000 sujets. Nous avons ensuite appliqué les méthodes d'analyse du fonctionnement différentiel à chaque échantillon et relevé les indices qui en découlent. À titre de rappel, ces indices sont le delta et le khi carré de Mantel-Haenszel (Δ_{MH} et X^2_{MH}), la différence de difficulté standardisée modifiée (DDSM) par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993, 1995) et, pour la méthode de régression logistique, le khi carré d'amélioration ($X^2_{amél.}$).

Le delta de Mantel-Haenszel (Δ_{MH}) et la différence de difficulté standardisée modifiée (DDSM), comme d'ailleurs la différence de difficulté standardisée régulière (DDS), constituent une mesure de l'ampleur des différences dans le fonctionnement des items. Ces indices distinguent les données du groupe de référence et celles du groupe focalisé et indiquent par un signe approprié, positif ou négatif, quel groupe est favorisé. Pour cette raison, on parle d'indices orientés. Le khi carré de Mantel-Haenszel (X^2_{MH}) et le khi carré d'amélioration pour la régression logistique ($X^2_{amél.}$) relèvent d'un test statistique dont les valeurs se distribuent selon la loi du khi carré, lorsque l'hypothèse nulle est vraie. Ils n'indiquent pas quel groupe est favorisé. C'est pourquoi on parle d'indices non orientés. Les deux groupes d'indices se comportent différemment. Les premiers ont des distributions à peu près symétriques. Les seconds ont des distributions asymétriques. De plus, l'augmentation du nombre de sujets dans les échantillons a un effet différent sur la grandeur et la variabilité des indices.

Pour examiner la distribution des indices en fonction de la grandeur des échantillons, nous avons déterminé pour chaque item la valeur minimale et la valeur maximale de chaque indice pour les 100 échantillons de même taille. Nous avons aussi déterminé la valeur de la médiane, du premier et du troisième quartile. Dans le cas des indices orientés, nous avons calculé la moyenne et l'écart-type. Les tableaux 10a, 10b, 10c et 10d résument les données sur la distribution des indices. Nous y indiquons, pour chaque item, la valeur minimale, la valeur maximale et la médiane des indices pour les 100 échantillons de même taille. On notera que les 15 items identifiés comme ayant un fonctionnement différentiel non négligeable dans la population apparaissent en premier. De plus, ils sont classés du plus grand au plus petit selon la grandeur des différences de difficulté standardisée (DDS) observées pour la population. Le déplacement et le classement de ces items ont pour but de faciliter la lecture et l'interprétation des données. Un relevé complet des données est fourni à l'annexe B (voir les tableaux 7a, 7b, 7c et 7d). Dans ce cas, nous avons laissé les items dans l'ordre où ils apparaissent dans le test.

Les médianes des deltas de Mantel-Haenszel et des différences de difficulté standardisée modifiée (Δ_{MH} et DDSM) varient généralement peu en fonction de la taille des échantillons. Par contre, la variabilité des indices décroît avec l'augmentation du nombre de sujets dans les échantillons, comme en témoignent la valeur minimale et la valeur maximale des indices. L'examen des moyennes, des écarts-types et des valeurs au premier et au troisième quartiles confirment cette double tendance. Le phénomène était prévisible. Un plus grand nombre de sujets accroît la précision de l'estimation des fréquences et des proportions de réponses correctes et, par ricochet, l'estimation de la mesure des différences dans le fonctionnement des items. Cette double tendance est plus manifeste pour les items qui ont un fonctionnement différentiel non négligeable dans la population.

Même si la variabilité des indices diminue avec l'accroissement du nombre de sujets dans les échantillons, la détection d'items qui ont un fonctionnement différentiel dans la population ne devrait pas varier beaucoup, si l'on se base sur la médiane et une valeur fixée a priori. Avec des échantillons de 250 sujets, tous les items peuvent être identifiés comme ayant un fonctionnement différentiel dans un échantillon. Le risque d'identifier à tort un item comme ayant un fonctionnement différentiel existe donc pour tous les items qui n'ont pas été retenus comme ayant un fonctionnement non négligeable dans la population. Avec des échantillons de 2000 sujets, le risque subsiste toujours, mais il est moins grand et limité à quelques items seulement.

Dans le cas du khi carré de Mantel-Haenszel et du khi carré d'amélioration pour la régression logistique (X^2_{MH} et $X^2_{amél.}$), la valeur des indices tend à croître avec l'augmentation du nombre de sujets dans les échantillons et la variabilité des indices s'en trouve amplifiée. Le phénomène existe pour tous les items, mais il est plus apparent pour les 15 items qui ont un fonctionnement différentiel non négligeable dans la population. De fait, l'augmentation de la grandeur du khi carré et celle de l'étendue de la distribution pour les 100 échantillons de même taille sont systématiques. Elles ne le sont pas pour les autres items. L'augmentation de la grandeur du khi carré et l'accroissement de l'étendue de la distribution en fonction de la taille des échantillons étaient prévisibles. L'augmentation du nombre de sujets est un moyen reconnu pour accroître la puissance d'un test statistique.

Quelle que soit la taille des échantillons, il y a une possibilité de ne pas détecter des items qui ont un fonctionnement différentiel dans la population. Il n'y a qu'une exception à la règle : les items 19 et 14 pour le khi carré de Mantel-Haenszel (X^2_{MH}) et les items 19 et 58 pour le khi carré d'amélioration ($X^2_{amél.}$). Avec des échantillons de 2000 sujets, la valeur minimale des indices est

supérieure à la valeur critique au seuil de 0,05. Par ailleurs, la médiane indique qu'avec des échantillons de 2000 sujets, on peut espérer un taux de détection d'au moins 50 pour cent pour chacun des items qui ont un fonctionnement différentiel non négligeable dans la population. Si les échantillons ne comptent que 1000 sujets, on peut espérer un taux de détection de 50 pour cent pour certains items, ceux dont les différences de fonctionnement sont les plus grandes, mais on ne peut espérer un tel taux pour tous les items. Avec des échantillons de 500 sujets, il n'y a que l'item 19 qui peut fournir un tel taux. Cet item est celui qui affiche la différence de fonctionnement la plus grande dans la population. Avec des échantillons de 250 sujets, on ne peut espérer un taux de détection de 50 pour cent, quels que soient les items et la grandeur des différences de fonctionnement des items dans la population.

Tableau 10a. Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour le delta de Mantel-Haenszel (N = 100 échantillons).

Items	Échantillons de 250 sujets			Échantillons de 500 sujets			Échantillons de 1000 sujets			Échantillons de 2000 sujets		
	Minimum	Maximum	Médiane	Minimum	Maximum	Médiane	Minimum	Maximum	Médiane	Minimum	Maximum	Médiane
19	-0,55	3,93	1,405	-0,24	2,8	1,475	0,53	2,83	1,42	0,84	2,18	1,425
17	-2,36	0,78	-1,02	-2,26	-0,01	-0,935	-1,99	-0,14	-1,07	-1,63	-0,09	-1,035
14	-1,97	3,08	1,12	-0,09	2,22	1,06	0,17	2,28	1,095	0,57	1,82	1,12
23	-2,08	1,22	-0,67	-1,93	0,35	-0,74	-1,68	0,13	-0,845	-1,33	-0,27	-0,805
58	-3,18	0,98	-1,185	-2,33	1,21	-1,195	-1,96	-0,25	-1,225	-1,9	-0,56	-1,2
52	-3,41	0,87	-0,945	-2,23	0,63	-0,825	-1,79	0,02	-0,795	-1,52	-0,27	-0,815
38	-2,96	2,12	-0,79	-2,49	0,03	-0,97	-2,06	-0,15	-0,875	-1,58	-0,37	-0,875
32	-2,31	0,6	-0,66	-2,28	0,66	-0,675	-1,77	0,36	-0,625	-1,35	-0,22	-0,68
29	-1,86	2,72	0,88	-0,7	2,14	0,87	-0,9	1,83	0,945	0,06	1,55	0,845
9	-0,81	2,45	0,8	-0,28	1,86	0,785	-0,13	1,55	0,685	0,08	1,33	0,69
7	-3,38	0,68	-0,87	-2,21	0,44	-0,76	-1,63	0,53	-0,635	-1,3	-0,09	-0,715
37	-1,23	2,4	0,72	-0,99	2,31	0,545	-0,3	1,69	0,62	0,25	1,21	0,64
4	-1,69	3,11	0,795	-1,2	2,07	0,77	-0,09	1,64	0,78	0,01	1,37	0,785
34	-2,32	1,65	-0,625	-2,05	0,4	-0,605	-1,62	0,2	-0,69	-1,29	-0,02	-0,665
18	-0,83	2,43	0,58	-0,64	1,79	0,46	-0,11	1,26	0,48	0,08	1,23	0,54
1	-2,44	3,03	0,14	-0,93	2,17	0,265	-0,88	1,27	0,315	-0,37	1,07	0,165
2	-3,2	1,39	-0,49	-2,48	0,92	-0,555	-1,44	0,45	-0,47	-1,39	0,14	-0,515
3	-2,53	2,56	0,23	-1,09	1,33	0,09	-0,72	1,01	0,035	-0,72	0,76	0,09
5	-1,37	2,28	0,415	-0,83	1,46	0,435	-0,59	1,35	0,31	-0,24	1,07	0,275
6	-2,2	1,7	-0,605	-2,39	0,74	-0,69	-1,54	0,43	-0,725	-1,12	-0,03	-0,6
8	-2,28	1,48	-0,275	-1,35	0,8	-0,23	-1,13	0,61	-0,275	-0,87	0,34	-0,295
10	-1,13	1,91	0,405	-0,83	2,25	0,44	-0,43	1,68	0,485	-0,08	1,15	0,45
11	-1,99	1,64	-0,205	-1,3	0,71	-0,14	-0,96	0,48	-0,285	-0,79	0,42	-0,27
12	-1,55	2,01	0,275	-0,77	1,51	0,315	-0,5	1,38	0,33	-0,42	0,89	0,315
13	-2,57	1,33	-0,41	-1,49	0,99	-0,305	-1,08	0,82	-0,375	-0,92	0,27	-0,335
15	-2,67	1,45	-0,47	-1,54	1,17	-0,38	-1,11	0,67	-0,455	-1,07	0,03	-0,43
16	-1,62	2,82	0,69	-0,75	2	0,655	-0,54	1,35	0,615	0	1,39	0,59
20	-2,09	1,82	-0,095	-1,1	0,9	0	-0,65	0,83	-0,03	-0,57	0,58	-0,035
21	-1,72	2,43	0,43	-0,55	1,88	0,35	-0,49	1,24	0,44	-0,45	0,89	0,39
22	-1,89	2,28	0,485	-0,74	1,47	0,45	-0,61	1,07	0,49	-0,15	1,03	0,49
24	-2,18	2,66	0,065	-0,85	1,18	0,155	-0,67	0,84	0,045	-0,77	0,52	-0,03
25	-1,44	2,43	0,225	-0,9	1,27	0,185	-0,59	0,99	0,16	-0,29	0,77	0,155
26	-1,74	2,34	0,125	-1,15	1,49	0,09	-0,87	0,76	0,015	-0,4	0,67	0,03
27	-1,73	1,84	-0,015	-1,32	1,56	-0,125	-0,72	0,7	-0,135	-0,68	0,43	-0,08
28	-2,77	1,8	-0,395	-1,35	1,02	-0,315	-1,05	0,53	-0,33	-0,73	0,4	-0,295
30	-1,85	1,59	0,16	-1,06	1,3	-0,02	-0,77	1,04	0,07	-0,52	0,78	0,1
31	-1,92	1,94	-0,125	-1,31	1,19	-0,175	-1,13	0,71	-0,29	-0,75	0,44	-0,24
33	-1,72	2,21	0,35	-1,12	1,48	0,155	-0,45	0,79	0,195	-0,37	0,62	0,18
35	-1,75	1,77	-0,305	-1,33	1,06	-0,18	-1,38	0,84	-0,08	-0,67	0,53	-0,08
36	-1,72	1,74	0,23	-1,14	1,51	0,28	-0,61	1,21	0,175	-0,6	0,63	0,12
39	-1,5	1,91	0,29	-1,03	1,38	0,29	-0,29	1,01	0,325	-0,43	0,78	0,29
40	-2,04	1,55	0,155	-0,87	1,2	0,15	-0,57	0,92	0,085	-0,5	0,61	0,01
41	-1,79	3,1	0,485	-0,63	2,05	0,58	-0,39	1,76	0,5	-0,14	1,25	0,565
42	-2,31	2,17	0,075	-0,88	1,04	0,145	-0,57	0,73	0,08	-0,38	0,62	0,09
43	-2,8	1,48	-0,295	-1,59	1,03	-0,215	-0,84	0,38	-0,225	-0,87	0,6	-0,21
44	-2,13	1,94	-0,07	-1,17	1,06	-0,195	-1,18	0,75	-0,1	-0,69	0,48	-0,14
45	-2,48	1,49	-0,31	-1,56	1,26	-0,19	-1,22	0,56	-0,2	-0,71	0,27	-0,22
46	-2,37	4,16	0,345	-1,57	2,14	0,475	-0,59	1,24	0,49	-0,42	1,36	0,43
47	-1,95	1,81	0,115	-1,24	1,43	0,035	-0,52	1,05	0,105	-0,34	0,87	0,17
48	-2,69	0,9	-0,535	-1,77	0,3	-0,61	-1,67	0,39	-0,57	-1,29	-0,04	-0,575
49	-1,95	2,23	0,205	-1,27	1,38	-0,04	-0,94	1,03	-0,01	-0,57	0,63	0,115
50	-1,31	1,98	0,11	-1,19	1,56	0,185	-0,85	1,05	0,245	-0,35	0,68	0,14
51	-1,32	2,72	0,46	-1,31	1,64	0,47	-0,59	1,46	0,35	-0,27	1,11	0,395
53	-2,46	2,58	-0,035	-1,91	1,29	-0,24	-1,16	0,87	-0,23	-0,88	0,35	-0,315
54	-1,91	2,18	0,165	-1,43	1,61	-0,03	-0,76	1,22	-0,005	-0,73	0,99	0,045
55	-2,61	2,89	-0,27	-1,76	1,09	-0,455	-1,24	0,7	-0,325	-0,86	0,34	-0,325
56	-3,01	2,13	-0,375	-2,1	1,32	-0,385	-1,43	1,02	-0,31	-1,12	0,55	-0,325
57	-3,12	3,64	0,58	-1,29	2,51	0,66	-0,3	1,99	0,59	-0,1	1,53	0,735
59	-3,06	3,02	0,33	-1,46	1,86	0,19	-0,67	1,3	0,28	-0,34	1,12	0,225
60	-1,95	3,06	0,22	-1,78	1,43	0,035	-0,8	1,08	-0,05	-1	0,71	0,055

Tableau 10b. Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour la différence de difficulté standardisée modifiée (N = 100 échantillons).

Items	Échantillons de 250 sujets			Échantillons de 500 sujets			Échantillons de 1000 sujets			Échantillons de 2000 sujets		
	Minimum	Maximum	Médiane	Minimum	Maximum	Médiane	Minimum	Maximum	Médiane	Minimum	Maximum	Médiane
19	-0,027	0,258	0,106	-0,002	0,222	0,108	0,027	0,214	0,105	0,058	0,15	0,1045
17	-0,205	0,101	-0,079	-0,22	0,015	-0,0855	-0,167	0,011	-0,088	-0,133	0	-0,0865
14	-0,072	0,26	0,075	-0,036	0,161	0,0755	0,009	0,148	0,075	0,027	0,117	0,0765
23	-0,226	0,142	-0,0775	-0,192	0,029	-0,074	-0,159	0,013	-0,079	-0,131	0,075	-0,0815
58	-0,157	0,037	-0,067	-0,14	0,015	-0,0635	-0,116	0,01	-0,065	-0,106	-0,019	-0,062
52	-0,223	0,083	-0,0795	-0,194	0,058	-0,066	-0,144	0,014	-0,067	-0,133	-0,021	-0,065
38	-0,184	0,057	-0,06	-0,153	0,02	-0,0685	-0,129	-0,013	-0,0645	-0,116	-0,006	-0,0615
32	-0,24	0,108	-0,046	-0,332	0,089	-0,0525	-0,153	0,039	-0,0485	-0,108	0,05	-0,049
29	-0,123	0,191	0,0515	-0,06	0,33	0,055	-0,003	0,115	0,0575	0	0,107	0,053
9	-0,06	0,2	0,0705	-0,009	0,161	0,0685	-0,008	0,128	0,058	0,013	0,12	0,0555
7	-0,215	0,138	-0,0795	-0,148	0,033	-0,0645	-0,139	0,04	-0,056	-0,103	0,057	-0,06
37	-0,066	0,23	0,065	-0,108	0,201	0,0395	-0,022	0,144	0,046	-0,002	0,105	0,0515
4	-0,093	0,179	0,051	-0,077	0,145	0,0525	-0,015	0,116	0,0545	-0,001	0,098	0,053
34	-0,233	0,119	-0,0365	-0,118	0,053	-0,0405	-0,12	0,03	-0,0435	-0,093	0,067	-0,041
18	-0,068	0,169	0,0505	-0,039	0,152	0,0425	-0,001	0,113	0,048	-0,003	0,097	0,05
1	-0,154	0,177	0,0155	-0,075	0,134	0,023	-0,18	0,083	0,016	-0,032	0,06	0,014
2	-0,278	0,102	-0,034	-0,174	0,101	-0,04	-0,11	0,038	-0,0405	-0,105	0,012	-0,04
3	-0,145	0,168	0,0275	-0,098	0,094	0,0135	-0,041	0,088	0,014	-0,038	0,071	0,0195
5	-0,101	0,167	0,052	-0,059	0,14	0,0445	-0,048	0,126	0,0295	-0,02	0,102	0,03
6	-0,172	0,114	-0,027	-0,165	0,048	-0,0465	-0,123	0,021	-0,055	-0,088	0,001	-0,044
8	-0,178	0,136	-0,032	-0,153	0,089	-0,0205	-0,098	0,055	-0,0215	-0,081	0,025	-0,0245
10	-0,107	0,202	0,0315	-0,109	0,187	0,0415	-0,036	0,131	0,046	-0,007	0,113	0,0455
11	-0,165	0,135	-0,0245	-0,12	0,059	-0,011	-0,098	0,05	-0,0255	-0,075	0,052	-0,022
12	-0,142	0,154	0,012	-0,079	0,148	0,0275	-0,045	0,101	0,0255	-0,035	0,063	0,02
13	-0,206	0,122	-0,026	-0,127	0,092	-0,021	-0,114	0,07	-0,028	-0,088	0,037	-0,026
15	-0,265	0,084	-0,045	-0,148	0,085	-0,038	-0,119	0,053	-0,0455	-0,088	0,003	-0,0445
16	-0,122	0,162	0,046	-0,051	0,127	0,0475	-0,041	0,107	0,045	-0,001	0,107	0,042
20	-0,218	0,174	-0,0055	-0,08	0,099	0,021	-0,048	0,096	0,0155	-0,049	0,088	0,013
21	-0,151	0,197	0,031	-0,075	0,182	0,0265	-0,049	0,127	0,0355	-0,039	0,099	0,03
22	-0,073	0,191	0,0355	-0,066	0,112	0,035	-0,051	0,099	0,0385	-0,026	0,069	0,0355
24	-0,152	0,162	0,008	-0,088	0,123	0,0055	-0,062	0,075	0,003	-0,06	0,042	-0,004
25	-0,151	0,193	0,028	-0,091	0,109	0,0145	-0,066	0,095	0,011	-0,038	0,071	0,011
26	-0,144	0,188	0,007	-0,102	0,128	0,0095	-0,084	0,078	-0,006	-0,047	0,053	-0,005
27	-0,15	0,149	-0,0145	-0,116	0,141	-0,0135	-0,078	0,056	-0,0165	-0,08	0,036	-0,014
28	-0,211	0,123	-0,027	-0,119	0,087	-0,0245	-0,088	0,055	-0,0205	-0,061	0,043	-0,0205
30	-0,14	0,143	0,011	-0,101	0,1	0	-0,066	0,081	0,002	-0,049	0,066	0,0085
31	-0,13	0,12	-0,011	-0,093	0,1	-0,0155	-0,079	0,069	-0,0185	-0,052	0,026	-0,012
33	-0,153	0,173	0,006	-0,101	0,18	0,0155	-0,048	0,074	0,0165	-0,035	0,067	0,019
35	-0,142	0,157	-0,024	-0,099	0,081	-0,0135	-0,117	0,074	-0,0025	-0,057	0,041	-0,005
36	-0,149	0,155	0,02	-0,116	0,126	0,0195	-0,065	0,096	0,0145	-0,069	0,058	0,007
39	-0,129	0,129	0,0205	-0,092	0,121	0,0235	-0,024	0,088	0,0205	-0,039	0,071	0,0215
40	-0,2	0,153	0,007	-0,086	0,1	0,0155	-0,33	0,085	0,0115	-0,051	0,059	0,0015
41	-0,117	0,132	0,0255	-0,04	0,119	0,0315	-0,029	0,1	0,0285	-0,017	0,08	0,03
42	-0,179	0,168	-0,002	-0,088	0,082	0,009	-0,059	0,054	0,0065	-0,042	0,057	0,004
43	-0,161	0,121	-0,025	-0,112	0,068	-0,0175	-0,071	0,04	-0,012	-0,054	0,039	-0,0165
44	-0,195	0,121	-0,0105	-0,121	0,07	-0,016	-0,089	0,067	-0,01	-0,05	0,044	-0,009
45	-0,176	0,098	-0,013	-0,092	0,077	-0,01	-0,076	0,048	-0,0145	-0,039	0,022	-0,011
46	-0,118	0,126	0,0185	-0,09	0,102	0,0195	-0,038	0,066	0,019	-0,018	0,075	0,02
47	-0,194	0,143	0,005	-0,118	0,084	-0,015	-0,058	0,061	-0,003	-0,038	0,051	0,002
48	-0,215	0,127	-0,0405	-0,129	0,044	-0,04	-0,11	0,078	-0,0315	-0,081	0,016	-0,033
49	-0,123	0,165	0,002	-0,094	0,093	-0,0005	-0,069	0,066	-0,008	-0,053	0,05	-0,0015
50	-0,146	0,176	0,01	-0,128	0,126	0,006	-0,07	0,102	0,017	-0,035	0,058	0,0105
51	-0,084	0,14	0,0215	-0,061	0,085	0,019	-0,03	0,07	0,0145	-0,015	0,047	0,0175
53	-0,104	0,106	0,0015	-0,121	0,079	-0,0115	-0,079	0,054	-0,0075	-0,055	0,036	-0,014
54	-0,119	0,129	0,0035	-0,114	0,061	-0,004	-0,054	0,052	-0,003	-0,032	0,047	-0,0015
55	-0,132	0,131	-0,018	-0,132	0,053	-0,0245	-0,077	0,05	-0,021	-0,054	0,029	-0,0155
56	-0,104	0,097	-0,012	-0,085	0,076	-0,0125	-0,083	0,054	-0,007	-0,055	0,04	-0,012
57	-0,078	0,102	0,015	-0,056	0,077	0,015	-0,022	0,063	0,017	-0,012	0,053	0,021
59	-0,33	0,13	0,0145	-0,066	0,083	0,008	-0,035	0,05	0,013	-0,016	0,057	0,009
60	-0,089	0,094	0,007	-0,075	0,056	-0,001	-0,036	0,039	0,0005	-0,035	0,032	0,002

Tableau 10c. Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour le khi carré de Mantel-Haenszel (N = 100 échantillons).

Items	Échantillons de 250 sujets			Échantillons de 500 sujets			Échantillons de 1000 sujets			Échantillons de 2000 sujets		
	Minimum	Maximum	Médiane	Minimum	Maximum	Médiane	Minimum	Maximum	Médiane	Minimum	Maximum	Médiane
19	0	17,71	2,57	0,03	25,99	7,23	1,86	52,05	14,195	10,86	65,97	29,805
17	0	8,83	1,59	0	21,36	3,36	0,13	33,89	10,295	0,12	48,66	19,915
14	0	11,9	1,52	0	15,16	3,09	0,14	31,09	7,515	4,38	41,87	16,86
23	0	7,13	0,915	0	15,23	2,115	0,02	23,95	5,965	1,17	30,23	11,385
58	0	10,56	1,22	0,01	15,16	3,86	0,31	24,13	8,6	3,75	42,72	17,765
52	0	12,9	1,115	0	17,3	2,41	0	24,9	4,815	1,12	39,3	10,855
38	0	11,74	0,74	0	18,6	2,975	0,1	28,69	4,85	1,8	35,71	10,925
32	0	11,95	0,59	0	20,14	1,87	0	27,25	3,505	0,85	33,45	8,515
29	0	8,14	0,88	0	13,23	2,07	0,05	19,33	5,295	0,02	31,84	9,29
9	0	8,27	0,725	0	11,15	2,04	0	17,63	3,545	0,08	27,68	7,435
7	0	13,01	0,92	0	13,21	1,695	0,04	17,57	2,595	0,09	24,34	7,01
37	0	10,47	0,825	0	20,77	1,175	0	23,56	3,315	1,05	25,58	7,145
4	0	9,65	0,74	0	10,23	1,46	0	16,23	3,345	0	22,77	7,23
34	0	6,81	0,61	0	15,33	1,125	0,01	19,4	3,415	0	26,58	6,99
18	0	7,84	0,51	0	10,16	0,78	0	11,94	1,8	0,09	25,52	4,835
1	0	8,98	0,185	0	10,71	0,29	0	8,66	0,6	0	11,72	0,615
2	0	11,92	0,275	0	15,81	0,93	0	11,48	1,155	0	23,33	3,11
3	0	5,27	0,175	0	4,07	0,295	0	4,95	0,28	0	6,43	0,32
5	0	8,24	0,335	0	7,47	0,575	0	13,09	0,765	0	18,09	1,155
6	0	5,43	0,575	0	15,52	1,185	0	12,38	3,03	0	15,64	4,36
8	0	9,4	0,295	0	7,45	0,285	0	10,02	0,705	0	13,21	1,64
10	0	6,42	0,335	0	19,87	0,785	0	20,65	1,68	0	21,29	3,185
11	0	7,12	0,35	0	7,05	0,19	0	8,07	0,865	0	11,13	1,265
12	0	5,19	0,335	0	8,09	0,725	0	14,34	0,93	0	12,55	1,52
13	0	10,5	0,6	0	8,79	0,48	0	10,93	1,275	0	16,16	2,095
15	0	11,56	0,455	0	8,39	0,63	0	9,29	1,53	0	18,2	2,91
16	0	7,98	0,64	0	11,09	1,245	0	12,52	2,52	0	27,96	5,015
20	0	7,84	0,18	0	5,08	0,21	0	6,45	0,32	0	6,58	0,335
21	0	8,84	0,425	0	14,18	0,565	0	13,64	1,8	0	14,79	2,81
22	0	6,42	0,35	0	8,08	0,795	0	8,83	1,84	0	17,61	3,84
24	0	8,59	0,165	0	4,84	0,285	0	4,93	0,25	0	8,97	0,42
25	0	7,95	0,285	0	5,69	0,325	0	8,56	0,455	0	10,53	0,5
26	0	8,05	0,195	0	7,71	0,475	0	5,94	0,37	0	7,67	0,345
27	0	5,3	0,245	0	9,25	0,275	0	4,44	0,25	0	8,3	0,315
28	0	12,06	0,39	0	5,97	0,46	0	7,8	0,74	0	7,99	1,46
30	0	5,7	0,285	0	6,19	0,2	0	7,67	0,265	0	9,66	0,84
31	0	5,04	0,28	0	5,32	0,33	0	9,23	0,645	0	8	0,79
33	0	7,56	0,33	0	8,22	0,545	0	5,5	0,46	0	6,97	0,645
35	0	6	0,4	0	6,45	0,32	0	15,65	0,225	0	7,59	0,395
36	0	4,52	0,36	0	8,88	0,445	0	11,59	0,43	0	6,19	0,415
39	0	5,65	0,35	0	7,06	0,69	0	7,84	0,755	0	9,91	1,395
40	0	6,51	0,28	0	5,87	0,27	0	6,86	0,34	0	6,82	0,37
41	0	8,82	0,31	0	11,22	0,97	0	17,99	1,47	0,01	20,7	4,285
42	0	7,25	0,235	0	3,4	0,315	0	4,33	0,29	0	5,82	0,365
43	0	8,97	0,485	0	8,54	0,315	0	4,79	0,44	0	10,87	0,865
44	0	6,89	0,31	0	5,07	0,375	0	10,55	0,66	0	7,84	0,565
45	0	9,15	0,23	0	6,82	0,36	0	10,85	0,52	0	7,44	0,755
46	0	8,47	0,255	0	10,97	0,485	0	7,49	1,215	0,05	21,13	2,015
47	0	7,32	0,245	0	6,25	0,345	0	8,26	0,47	0	12,35	0,585
48	0	10,22	0,375	0	10,38	1,06	0	19,62	2,305	0,01	24,69	4,795
49	0	7,14	0,38	0	5,76	0,39	0	7,93	0,37	0	5,94	0,48
50	0	6,43	0,305	0	8,98	0,41	0	8,73	0,56	0	8,26	0,395
51	0	6	0,26	0	6,53	0,59	0	10,39	0,74	0	13,81	1,65
53	0	7,87	0,25	0	9,7	0,31	0	8,58	0,59	0	10,58	1,27
54	0	5,69	0,32	0	6	0,175	0	7,66	0,36	0	10,45	0,495
55	0	9,15	0,265	0	10,77	0,595	0	10,12	0,755	0	9,72	1,375
56	0	4,11	0,275	0	7,73	0,465	0	10,47	0,575	0	11,36	1,205
57	0	6,3	0,25	0	8,78	0,65	0	15,78	1,235	0,01	18,12	4,29
59	0	7,93	0,28	0	7,58	0,31	0	8,19	0,405	0	14,25	0,5
60	0	6,92	0,16	0	4,9	0,19	0	3,77	0,235	0	7,25	0,43

Tableau 10d. Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour le khi carré d'amlioration et la régression logistique (N = 100 échantillons).

Items	Échantillons de 250 sujets			Échantillons de 500 sujets			Échantillons de 1000 sujets			Échantillons de 2000 sujets		
	Minimum	Maximum	Médiane	Minimum	Maximum	Médiane	Minimum	Maximum	Médiane	Minimum	Maximum	Médiane
19	0,068	30,298	4,586	0,001	31,485	9,412	1,82	59,197	16,253	10,326	70,605	33,319
17	0,021	15,35	3,893	0,213	29,205	5,5635	0,872	41,883	13,818	2,352	51,097	22,4745
14	0,117	17,417	3,538	0,017	19,292	5,901	0,18	32,096	9,795	1,81	47,817	18,7845
23	0,047	13,799	2,708	0,072	18,515	4,504	0,746	26,766	8,0675	1,712	32,801	13,967
58	0,027	17,252	4,156	0,642	19,075	6,62	0,732	36,903	11,402	7,805	51,523	22,1615
52	0,009	15,48	3,0125	0,015	19,484	4,64	0,225	27,975	7,209	2,447	41,258	14,3745
38	0,014	14,731	2,673	0,062	20,627	5,25	0,463	27,557	7,3075	3,854	40,071	12,4905
32	0,007	16,465	2,2235	0,071	21,061	3,765	0,378	28,209	6,922	0,818	37,22	11,8615
29	0,102	32,939	2,4085	0,003	19,359	4,5285	0,617	22,26	7,4035	0,299	34,76	11,579
9	0,066	13,648	2,2495	0,112	14,229	3,4965	0,007	19,248	5,005	0,44	31,381	8,673
7	0,032	14,378	2,6255	0,078	17,356	3,6575	0,089	17,094	4,35	1,805	24,538	8,785
37	0,049	13,837	2,1255	0,139	20,846	2,537	0,094	24,233	4,945	1,679	25,039	8,4385
4	0,023	14,739	2,1945	0,205	14,649	3,422	0,112	23,017	5,3785	1,22	33,768	10,3565
34	0,015	12,413	2,0785	0,028	17,391	2,9495	0,309	20,382	6,103	0,303	31,192	10,7105
18	0,066	9,932	1,455	0,001	17,846	2,2855	0,094	13,449	3,342	0,331	30,12	6,684
1	0,006	12,132	1,543	0,038	20,067	2,126	0,01	14,718	2,1815	0,112	25,114	2,384
2	0,016	16,285	1,802	0	18,605	2,127	0,095	18,167	3,7655	0,281	25,867	6,393
3	0,052	11,647	2,208	0,014	9,194	1,515	0,152	16,721	2,2435	0,125	17,384	3,194
5	0,005	13,677	1,5925	0,003	9,262	2,1915	0,005	14,821	1,93	0,033	20,568	2,972
6	0,001	15,367	1,9465	0,011	20,74	2,5855	0,042	16,7	4,688	0,224	18,981	5,8775
8	0,018	10,938	1,803	0,035	12,793	1,2125	0,112	9,55	1,667	0	14,341	2,3685
10	0,02	13,557	1,8575	0,062	17,5	2,4005	0,027	20,621	3,205	0,085	21,646	5,048
11	0,007	13,373	1,215	0,006	8,939	1,578	0,001	14,442	2,2295	0,01	17,043	2,502
12	0,001	9,585	1,311	0,012	13,668	1,9275	0,004	17,224	1,9955	0,018	16,112	2,5075
13	0,066	11,813	1,678	0,034	11,135	1,8225	0,055	15,755	3,1265	0,422	20,746	4,6525
15	0,005	16,675	2,3305	0,083	15,682	2,4225	0,025	14,78	4,1735	0,478	20,139	7,438
16	0,017	9,795	1,696	0,058	22,743	2,768	0,232	19,774	4,271	0,025	31,197	6,791
20	0,013	12,796	1,581	0,037	17,903	1,8115	0,048	18,475	2,9745	0,233	27,09	4,9995
21	0,001	11,415	1,3595	0,014	17,012	1,997	0,005	15,436	2,794	0,079	17,436	4,4445
22	0,017	12,346	2,092	0,003	13,786	1,976	0,05	11,633	3,19	0,025	28,64	5,0945
24	0,028	9,746	1,5525	0,009	8,867	1,2225	0,023	9,226	1,3145	0,005	10,485	1,3085
25	0,052	12,34	1,128	0,026	8,93	1,6215	0,018	9,601	1,3135	0,004	12,885	1,4815
26	0,024	10,422	1,3645	0,047	10,339	1,4955	0,005	9,076	1,1865	0,006	7,874	1,196
27	0,028	11,103	1,888	0,03	11,436	1,668	0,023	9,481	1,425	0,007	13,019	2,354
28	0,017	14,235	1,7095	0,007	9,381	1,9725	0,001	9,639	2,1705	0,005	13,329	3,1
30	0,014	7,589	1,4105	0,017	11,851	1,401	0,002	8,856	0,9555	0,056	12,26	1,673
31	0,009	9,84	1,34	0,007	10,526	1,716	0,064	10,618	1,7675	0	12,745	2,281
33	0,015	13,871	1,836	0,043	26,528	2,0755	0,085	12,642	2,8275	0,132	25,236	4,8205
35	0,011	13,722	1,794	0,021	12,799	1,3635	0,022	14,498	1,5635	0,016	11,834	2,228
36	0,006	9,874	1,315	0,029	10,672	1,2965	0,025	12,019	1,159	0,046	11,787	1,2355
39	0,008	6,103	1,41	0,043	8,314	1,606	0,009	9,189	1,894	0,014	106,181	2,074
40	0,008	12,604	1,6395	0,029	12,625	1,07	0,059	8,692	1,3405	0,073	7,151	1,475
41	0,068	13,466	1,543	0,011	12,709	2,158	0,043	21,822	3,2925	0,149	25,688	5,795
42	0,017	13,239	1,556	0,005	7,589	1,3775	0,014	7,555	1,8355	0,041	12,839	1,8475
43	0,048	11,844	1,619	0,009	8,203	1,5	0,013	7,902	1,45	0,05	14,977	2,6085
44	0,014	9,09	1,0105	0,011	9,829	1,563	0,052	12,877	1,817	0,04	12,217	1,852
45	0,003	14,745	1,2085	0,016	7,691	1,551	0,008	16,363	1,9675	0,012	13,142	3,126
46	0,01	18,42	1,5305	0,055	15,077	1,853	0,056	10,565	2,0085	0,232	29,814	3,6015
47	0,003	10,529	1,0045	0,008	9,246	1,6685	0,008	11,773	1,527	0,041	16,318	1,695
48	0,004	14,567	2,1825	0,055	14,267	3,0125	0,027	18,08	4,415	0,331	28,734	7,7355
49	0,002	10,711	1,3845	0,001	7,254	1,201	0,026	21,042	1,258	0,02	7,671	1,9275
50	0,001	11,289	1,1265	0,019	11,153	1,3475	0,027	16,169	1,26	0,029	9,146	1,2525
51	0,02	15,505	1,5155	0,054	9,374	1,7935	0,017	11,35	1,7535	0,048	16,255	2,956
53	0,006	10,073	1,7995	0,033	11,602	2,254	0,064	16,509	2,7725	0,46	17,255	4,3395
54	0,005	12,026	1,6005	0,004	11,542	1,068	0,003	7,965	1,245	0,004	13,578	1,507
55	0,001	18,411	1,845	0,033	15,433	2,5185	0,072	14,718	4,222	0,073	28,12	7,141
56	0,07	9,999	1,8685	0	13,981	2,708	0,293	22,987	5,187	0,656	24,187	9,43
57	0,032	11,259	1,647	0,099	19,655	2,252	0,001	14,876	2,7965	0,002	21,039	5,2795
59	0,002	16,804	1,488	0,012	11,284	1,427	0,053	9,858	1,5785	0,008	17,771	1,9165
60	0,082	14,527	1,599	0,091	13,845	2,138	0,027	19,545	2,443	0,198	22,881	4,31

Étude de la fidélité et de la validité des méthodes

Afin de répondre aux questions de recherche, nous avons déterminé la stabilité des indices pour les échantillons de même taille et étudié la validité et la fidélité des décisions par rapport aux indices de fonctionnement différentiel estimés pour la population. Stabilité et validité ont donc été étudiées pour des échantillons de 250, de 500, de 1000 et de 2000 sujets. À titre de rappel, les échantillons de 250 sujets comptent 125 personnes par groupe ; les échantillons de 500 sujets, 250 personnes par groupe ; ceux de 1000 sujets, 500 personnes par groupe ; et ceux de 2000 sujets, 1000 personnes par groupe. Nous examinerons d'abord la stabilité des indices et ensuite leur validité.

La stabilité des indices

L'évaluation de la stabilité des indices de fonctionnement différentiel repose ici sur la corrélation entre les indices de même nature pour les 100 échantillons de même taille pris deux à deux. Notre but est de voir jusqu'à quel point il est possible de détecter les mêmes items quels que soient les échantillons. Pour ce faire, nous avons assumé que les indices de fonctionnement différentiel devaient classer les items de façon assez semblable dans tous les échantillons de sorte que l'on puisse détecter les mêmes items peu importe les échantillons, pourvu que la grandeur des échantillons soit la même. Pour les indices de fonctionnement différentiel qui s'appuient sur une mesure de l'ampleur des différences dans le fonctionnement des items (Δ_{MH} et DDSM), nous avons vérifié l'hypothèse à l'aide du coefficient de corrélation de Pearson. Pour les indices qui reposent sur des khis carrés (X_{MH} et $X_{amél.}$), nous avons eu recours au coefficient de corrélation de rangs de Spearman. Le tableau 11 résume les données recueillies pour chaque type d'indices et

chaque grandeur d'échantillon. Nous y indiquons la moyenne, la médiane et l'étendue des coefficients de stabilité pour les 100 échantillons de même taille pris deux à deux. Nous examinerons d'abord les coefficients de stabilité des indices basés sur une mesure de l'ampleur des différences dans le fonctionnement des items. Nous examinerons les autres ensuite. Nous terminerons par une appréciation globale de la stabilité des indices.

Qu'il s'agisse du delta de Mantel-Haenszel (Δ_{MH}) ou de la différence de difficulté standardisée modifiée (DDSM), les coefficients de stabilité des échantillons de 250 sujets sont généralement faibles, et leur variabilité est énorme. Pour les deltas de Mantel-Haenszel (Δ_{MH}), ils varient de -0,295 à 0,667 avec une moyenne de 0,296. Un peu plus de la moitié sont positifs et significatifs au seuil de 0,05. Les autres sont ou négatifs et significatifs ou trop faibles pour être significatifs. Ils vont de 0,132 à 0,765 avec une moyenne de 0,496 pour les échantillons de 500 sujets. Presque tous sont significatifs au seuil de 0,05 et plus de la moitié le sont au seuil de 0,01. Toutefois, ce n'est que pour les échantillons de 1000 et de 2000 sujets qu'ils atteignent généralement un seuil acceptable. Pour les échantillons de 1000 sujets, les coefficients varient de 0,410 à 0,864 avec une moyenne de 0,672. Pour les échantillons de 2000 sujets, ils vont de 0,656 à 0,907 avec une moyenne de 0,810. Dans les deux cas, tous sont significatifs au seuil de 0,01.

Tableau 11. Stabilité des indices de fonctionnement différentiel d'un échantillon à l'autre pour les échantillons de taille réduite (N = 4 950 comparaisons pour chaque grandeur d'échantillon).

Indices		Échantillons			
		250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}	**Moyenne**	**0,296***	**0,496***	**0,672***	**0,810***
	Étendue	de -0,295* à 0,667*	de 0,132 à 0,765*	de 0,410* à 0,864*	de 0,656* à 0,907*
	Médiane	0,304*	0,501*	0,677*	0,812*
DDSM	**Moyenne**	**0,281***	**0,473***	**0,634***	**0,776***
	Étendue	de -0,198 à 0,659*	de 0,037 à 0,768*	de 0,325* à 0,850*	de 0,534* à 0,898*
	Médiane	0,290*	0,478*	0,641*	0,781*
X^2_{MH}	**Moyenne**	**0,052**	**0,175**	**0,354***	**0,539***
	Étendue	de -0,407* à 0,519*	de -0,266* à 0,578*	de -0,135 à 0,679*	de 0,206 à 0,770*
	Médiane	0,051	0,178	0,364*	0,541*
$X^2_{amél.}$	**Moyenne**	**0,051**	**0,160**	**0,338***	**0,493***
	Étendue	de -0,390* à 0,511*	de -0,264* à 0,624*	de -0,098 à 0,663*	de 0,167 à 0,792*
	Médiane	0,059	0,167	0,343*	0,497*

* Le coefficient de corrélation est significatif au seuil de 0,05.

** Les coefficients de stabilité des deltas de Mantel-Haenszel (Δ_{MH}) et ceux des différences de difficulté standardisée modifiée (DDSM) sont des coefficients de corrélation de Pearson. Les coefficients de stabilité des khis carrés de Mantel-Haenszel (X^2_{MH}) et ceux des khis carrés d'amélioration ($X^2_{amél.}$) sont des coefficients de corrélation de rangs de Spearman.

Les coefficients de stabilité des différences de difficulté standardisée modifiée (DDSM) suivent de près les coefficients de stabilité des deltas de Mantel-Haenszel, mais ils sont généralement plus faibles. La différence entre les coefficients de stabilité moyens ou médians n'est jamais très grande pour les échantillons de même grandeur. Pour les échantillons de 500, de 1000 et de 2000 sujets, l'étendue de la distribution est plus grande. La valeur minimale de la distribution est nettement plus petite tandis que la valeur maximale l'est à peine. Pour les échantillons de 250 sujets, la situation est inversée. L'étendue de la distribution est plus petite et la valeur minimale est plus grande. À cette différence s'ajoutent, pour les échantillons de 1000 sujets, quelques coefficients significatifs au seuil de 0,05, mais non au seuil de 0,01. Dans le cas des deltas de Mantel-Haenszel, tous sont significatifs au seuil de 0,01.

La stabilité des indices de fonctionnement différentiel basés sur le khi carré de Mantel-Haenszel (X^2_{MH}) et le khi carré d'amélioration pour la régression logistique ($X^2_{amél.}$) paraît plus faible que celle des indices basés sur une mesure de l'ampleur des différences dans le fonctionnement des items (Δ_{MH} et DDSM). Toutefois, leur faiblesse est plus apparente que réelle. Ces coefficients résultent de corrélations de rangs. Dans le cas des indices orientés (Δ_{MH} et DDSM), il s'agit de corrélations de Pearson. L'écart entre les khis carrés des items qui ont un fonctionnement différentiel et les khis carrés des items qui n'en ont pas est généralement plus grand que celui qui sépare les items sans fonctionnement différentiel. Lorsqu'on transforme les khis carrés en rangs, on attribue à chaque item un rang différent sans tenir compte de la grandeur de l'écart qui sépare les khis carrés. Par la suite, lorsqu'on calcule la corrélation de rangs, on accorde la même importance à toutes les différences de rangs. Il peut en résulter une sous-estimation de la corrélation.

Plus important encore, les indices orientés distinguent les items qui favorisent les hommes de ceux qui favorisent les femmes en leur attribuant un signe positif ou négatif selon le groupe qui est favorisé. Par conséquent, les items sont répartis de part et d'autre de la moyenne. Les indices non orientés ne font pas cette distinction. Les indices qui favorisent les hommes et ceux qui favorisent les femmes s'entremêlent. Il en résulte une plus grande possibilité de variation dans le classement des items. Pour comparer les coefficients de stabilité des deux types d'indices, orientés et non orientés, il faudrait transformer les indices orientés en indices non orientés en utilisant les valeurs absolues, puis transformer celles-ci en rangs.

La stabilité des khis carrés de Mantel-Haenszel (X^2_{MH}) et celle des khis carrés d'amélioration pour la régression logistique ($X^2_{amél.}$) se ressemblent. Les deux types d'indices révèlent des tendances semblables à celles observées pour les indices orientés. Dans les deux cas, la stabilité des indices tend à croître avec l'augmentation du nombre de sujets dans les échantillons et la variabilité des indices tend à diminuer.

Les coefficients de stabilité des khis carrés de Mantel-Haenszel (X^2_{MH}) varient de -0,407 à 0,519 avec une moyenne de 0,052 pour les échantillons de 250 sujets. Mis à part les cas extrêmes, la plupart ne sont pas significatifs au seuil de 0,05. Pour les échantillons de 500 sujets, ils vont de -0,266 à 0,578 avec une moyenne de 0,175. Moins de la moitié sont significatifs au seuil de 0,05. Ils varient de -0,135 à 0,679 avec une moyenne de 0,354 pour les échantillons de 1000 sujets. La majorité sont significatifs non seulement au seuil de 0,05, mais au seuil de 0,01. Toutefois, ce n'est qu'avec les échantillons de 2000 sujets qu'ils atteignent un seuil minimalement acceptable dans la moitié des cas. Pour les échantillons de cette grandeur, les

coefficients vont de 0,206 à 0,770 avec une moyenne de 0,539. Presque tous sont significatifs au seuil de 0,05. La majorité le sont au seuil de 0,01.

Les coefficients de stabilité des khis carrés d'amélioration ($X^2_{amél}$) affichent des valeurs semblables à celles observées pour les khis carrés de Mantel-Haenszel (X^2_{MH}). Comme avec ces derniers, on constate que des échantillons de 250 sujets ne peuvent classer les items de façon semblable d'un échantillon à un autre de même taille, pour ce qui est des différences de fonctionnement des items. De fait, ce n'est qu'avec des échantillons de 2000 sujets que l'on peut espérer une stabilité minimalement acceptable dans le classement des items, quels que soient les échantillons, pourvu que les échantillons aient la même grandeur.

À partir des coefficients de stabilité observés, il appert que les indices basés sur les khis carrés classent les items différemment d'un échantillon à un autre de même taille. Les variations dans le classement des items sont d'autant plus grandes que la taille des échantillons est plus petite. Avec des échantillons de 250 sujets, il est plus que douteux que l'on puisse détecter les mêmes items dans deux échantillons de même taille. Pour les échantillons plus grands, il est difficile de porter un jugement à partir des données corrélationnelles observées. Il y a des variations dans le classement des items, mais on ne sait pas quels items sont touchés. Même si, pris globalement, le classement des items est peu stable, il est possible que les items dont les différences de fonctionnement sont les plus grandes aient tendance à se classer les premiers. Dans une telle éventualité, les items qui ont un fonctionnement non négligeable dans la population pourraient être détectés quels que soient les échantillons pourvu que les échantillons aient la même grandeur. Cette question met en cause la validité des items détectés par rapport aux items retenus comme ayant un fonctionnement non négligeable dans la population et la constance des

décisions dans les échantillons eu égard au critère de décision retenu. Elle sera étudiée plus loin, lors de l'examen de la validité et de la fidélité des décisions.

Pour l'instant, nous retenons que, quelle que soit la méthode, la stabilité des indices tend à croître avec l'augmentation du nombre de sujets dans les échantillons. Parallèlement, il y a diminution de la variabilité des coefficients de stabilité. Le rythme varie en fonction des méthodes. La figure 1, à l'annexe B, illustre les deux tendances. L'augmentation des coefficients de stabilité est attestée par l'augmentation des coefficients moyens de stabilité ; la diminution de la variabilité, par la diminution de l'étendue des distributions. Malgré cette double tendance, il y a beaucoup de chevauchement entre les distributions de coefficients de stabilité, d'une grandeur d'échantillon à l'autre. Nous avons vérifié si la différence entre les coefficients de stabilité moyens ou médians des indices de même nature était significative au seuil de 0,05 pour des échantillons de taille différente. Les échantillons ayant été prélevés de façon indépendante, nous avons postulé qu'il n'y avait pas de relation entre les coefficients. Quels que soient les indices (Δ_{MH}, DDSM, X^2_{MH} et $X^2_{amél.}$), la différence entre les coefficients de stabilité moyens ou médians s'avère non significative pour une grandeur d'échantillon immédiatement supérieure. Par contre, elle l'est lorsqu'on oppose les coefficients de stabilité moyens ou médians des échantillons de 250 sujets à ceux des échantillons de 1000 ou de 2000 sujets et ceux des échantillons de 500 sujets aux coefficients moyens ou médians des échantillons de 2000 sujets.

Nous avons aussi vérifié s'il y avait une différence significative entre les coefficients de stabilité moyens et médians de nature différente pour les échantillons de même taille. Parce que les sujets et les items sont les mêmes et que seul le type d'indices change, nous avons appliqué le test de Williams pour coefficients non indépendants. La vérification a porté sur les coefficients des

indices de difficulté standardisée modifiée (DDSM) par rapport aux coefficients des deltas de Mantel-Haenszel (Δ_{MH}). La différence n'est pas significative au seuil de 0,05.

Compte tenu des données recueillies, il appert que les résultats des analyses de fonctionnement différentiel effectuées avec de petits échantillons sont très dépendants de la taille des échantillons utilisés et sujets à d'importantes fluctuations d'échantillonnage. Peu importe la méthode et le type d'indices, des échantillons de 250 sujets fournissent des indices généralement peu stables et sujets à des fluctuations d'échantillonnage considérables. Ce n'est qu'avec des échantillons de 2000 sujets que la stabilité des khis carrés de Mantel-Haenszel et des khis carrés d'amélioration atteint un seuil généralement acceptable. Les deltas de Mantel-Haenszel et les différences de difficulté standardisée modifiée y parviennent généralement avec des échantillons de 1000 sujets. Toutefois, même si les coefficients de stabilité des indices orientés sont plus grands que ceux des indices non orientés, ces derniers pourraient s'avérer aussi fiables en pratique, quand vient le temps d'identifier les items qui ont un fonctionnement différentiel non négligeable. En ce qui concerne la méthode de standardisation modifiée, il appert que celle-ci fournisse des indices dont la stabilité est un peu moins grande que celle des deltas de Mantel-Haenszel. Toutefois, la différence est minime. Par ailleurs, mis à part le cas des échantillons de 250 sujets, la variabilité des coefficients de stabilité est toujours plus grande.

La validité et la fidélité des décisions

Pour étudier la validité et la fidélité des décisions, nous avons compté le nombre de fois que chaque item est détecté pour les 100 échantillons de même taille. Nous avons aussi estimé la corrélation entre les items détectés dans les échantillons et les items classés comme ayant un fonctionnement non négligeable dans la population. Enfin, nous avons déterminé le nombre de décisions correctes pour l'ensemble du test. Ces études ont été effectuées en fonction du critère de décision retenu pour déterminer quels items ont un fonctionnement différentiel dans les échantillons de taille réduite. Pour la présente étude, nous avons considéré quatre critères de décision.

Les trois premiers critères utilisent des centiles. L'utilisation de centiles repose sur l'hypothèse que les méthodes peuvent classer les items de façon assez semblable pour les items qui accusent les différences de fonctionnement les plus grandes dans la population, et ce, même si, pris globalement, il y a passablement de divergence dans le classement des items. Les critères retenus visent à identifier cinq, 10 ou 20 pour cent des items qui composent le test. Il s'agit des centiles $C_{2,5}$ et $C_{97,5}$ ou du centile C_5 pour la détection de cinq pour cent des items, des centiles C_5 et C_{95} ou du centile C_{90} pour la détection de 10 pour cent des items et des centiles C_{10} et C_{90} ou du centile C_{80} pour la détection de 20 pour cent des items. Nous utilisons deux centiles pour les indices orientés (Δ_{MH} et DDSM) et un seul centile pour les indices non orientés (X^2_{MH} et $X^2_{amél.}$).

Le quatrième critère se rapporte aux pratiques en usage. Pour la méthode du delta de Mantel-Haenszel (Δ_{MH}), nous utilisons la valeur recommandée par Holland et Thayer (1988), soit une valeur égale ou supérieure à l'unité en valeur absolue. Pour la méthode du khi carré de Mantel-

Haenszel (X^2_{MH}), il s'agit de la valeur critique du khi carré au seuil de 0,05 à un degré de liberté, soit 3,841. Pour la méthode de régression logistique($X^2_{amél.}$), il s'agit de la valeur du khi carré au seuil de 0,05 à deux degrés de liberté, soit la valeur de 5,991. Pour la méthode de standardisation modifiée (DDSM), il n'y a pas de valeur seuil recommandée ou utilisée à grande échelle. La méthode ayant été conçue pour remplacer la méthode de standardisation avec des échantillons de taille réduite, nous avons d'abord pensé utiliser les valeurs recommandées pour cette dernière, soit 0,05 ou 0,10 abstraction faite du signe qui indique quel groupe est favorisé. L'examen des données a révélé qu'un seuil de 0,05 en valeur absolue était trop faible. Trop d'items étaient identifiés à tort comme ayant un fonctionnement différentiel dans les petits échantillons. Nous avons donc décidé de nous baser sur les écarts-types pour fixer la valeur seuil. Pour ce faire, nous avons calculé les indices moyens pour les 100 échantillons de 250, de 500, de 1000 et de 2000 sujets et l'écart-type de chaque distribution. Nous avons ensuite déterminé la valeur médiane des écarts-types pour les quatre distributions d'indices moyens et la distribution des indices pour les 40 000 sujets représentant la population. Les écarts-types ainsi obtenus sont 0,037852 ; 0,038558 ; 0,039377 ; 0,038535 pour les quatre séries d'indices moyens et 0,037118 pour les indices de la population. La médiane est donc de 0,038535. La valeur d'indice à 1,96 écart-type est alors de 0,075529. Les valeurs d'indices retenues ne comptant que trois chiffres après la virgule, nous utilisons la valeur de 0,075. Fait à noter, en appliquant la même procédure aux deltas de Mantel-Haenszel (Δ_{MH}), nous obtenons une valeur proche de l'unité. Pour les quatre distributions d'indices moyens, les écarts-types sont de 0,52861, 0,53032, 0,53666 et 0,54412. Pour la population, l'écart-type de la distribution des indices est de 0,52056. La médiane des écarts-types est alors de 0,53032 et la valeur d'indice à 1,96 écart-type est de 1,03942 ou 1,039 si l'on ne retient que trois chiffres après la virgule.

Nous considérerons d'abord le nombre de fois que chaque item est détecté pour chaque grandeur d'échantillon et chaque critère décisionnel, ce qui pourra donner un aperçu de la fidélité des décisions pour chaque item. Nous examinerons ensuite les corrélations entre les items classés comme ayant un fonctionnement différentiel dans les échantillons et les items identifiés comme ayant un fonctionnement différentiel non négligeable dans la population. Nous terminerons par l'examen des données relatives au nombre de décisions correctes pour les 100 échantillons de même taille.

Fréquence de détection de chaque item

Les tableaux 12a, 12b, 12c et 12d font état du nombre de fois qu'un item est détecté pour chaque item, chaque méthode, chaque grandeur d'échantillon et chaque critère de décision. Les 15 items qui se sont avérés avoir un fonctionnement différentiel non négligeable dans la population figurent en premier. Leur déplacement a pour but de faciliter la lecture et l'interprétation des données. Comme dans le cas des tableaux 10, nous les avons classés du plus grand au plus petit en fonction des différences de fonctionnement estimées pour la population à partir des différences de difficulté standardisée (DDS). Les autres items n'ont pas été classés de la sorte, pour éviter les risques d'erreur. Quels que soient le critère et la taille des échantillons, rares sont les items qui présentent des fréquences de détection de 100 pour les items avec un fonctionnement différentiel non négligeable dans la population.

Tableau 12a. Nombre de fois qu'un item est identifié comme ayant un fonctionnement différentiel dans les échantillons de même taille avec le centile C95 ou les centiles C2,5 et C97,5 comme critère de décision (N = 100 échantillons).

Items	Méthode du delta de Mantel-Haenszel				Méthode de standardisation modifiée			
	250 sujets	500 sujets	1000 sujets	2000 sujets	250 sujets	500 sujets	1000 sujets	2000 sujets
19	30	55	73	92	29	57	79	96
17	17	24	38	51	20	36	46	59
14	26	29	39	52	18	26	33	46
23	5	11	7	8	22	25	37	51
58	28	40	60	67	9	13	8	12
52	9	14	19	18	14	23	26	25
38	15	19	14	26	6	11	13	16
32	6	12	11	5	7	18	12	9
29	15	15	24	17	9	10	10	9
9	8	19	11	4	11	22	14	11
7	15	7	10	3	14	12	16	16
37	9	3	3	5	17	8	10	8
4	9	12	12	8	8	8	12	10
34	4	6	7	9	9	7	7	3
18	3	4	2	2	7	4	6	7
1	7	3	2	1	7	3	2	0
2	8	13	4	3	6	7	5	2
3	6	1	0	0	8	2	1	0
5	6	1	0	1	8	8	3	3
6	8	12	6	4	8	11	4	2
8	2	1	1	0	9	1	0	1
10	2	5	3	2	8	9	13	5
11	1	0	0	0	4	2	1	0
12	3	2	1	0	2	2	1	0
13	1	2	0	0	8	5	7	3
15	13	3	1	1	16	6	6	2
16	3	6	6	2	9	9	6	5
20	2	0	0	0	6	0	1	1
21	2	4	3	0	6	7	7	2
22	6	0	0	0	5	2	1	0
24	2	0	0	0	3	2	0	0
25	5	0	1	0	9	2	3	1
26	3	1	0	0	5	2	1	0
27	1	0	0	0	7	4	0	0
28	7	4	0	0	6	3	0	0
30	1	0	0	0	2	0	0	0
31	5	0	1	0	3	1	0	0
33	3	2	0	0	10	5	0	0
35	2	0	0	0	4	1	2	0
36	1	3	0	0	4	5	0	0
39	2	2	0	0	2	1	0	0
40	1	0	0	0	5	1	1	0
41	5	10	5	3	2	3	1	1
42	5	0	0	0	6	0	0	0
43	2	2	0	0	3	1	0	0
44	2	0	1	0	4	1	0	0
45	4	1	2	0	3	1	1	0
46	6	6	0	0	0	1	0	0
47	0	1	0	0	4	2	0	0
48	7	8	10	4	8	6	4	1
49	2	1	0	0	1	1	1	0
50	1	2	0	0	8	4	1	0
51	4	4	4	0	1	0	0	0
53	4	4	1	0	0	2	0	0
54	6	2	1	0	1	1	0	0
55	7	6	3	0	1	2	0	0
56	6	6	4	4	0	0	1	0
57	14	11	10	12	0	0	0	0
59	12	4	1	1	1	1	0	0
60	11	3	0	0	0	0	0	0

Tableau 12a. (suite) Nombre de fois qu'un item est identifié comme ayant un fonctionnement différentiel dans les échantillons de même taille avec le centile C95 ou les centiles C2,5 et C97,5 comme critère de décision (N = 100 échantillons).

Items	Méthode du khi carré de Mantel-Haenszel				Méthode de régression logistique			
	250 sujets	500 sujets	1000 sujets	2000 sujets	250 sujets	500 sujets	1000 sujets	2000 sujets
19	24	51	71	91	18	48	64	89
17	19	28	45	54	20	19	39	49
14	19	21	29	33	14	22	28	25
23	11	12	9	12	7	12	14	10
58	17	21	33	40	16	21	35	42
52	10	14	17	12	11	13	16	18
38	11	14	12	16	5	12	11	14
32	8	17	11	7	7	17	14	13
29	11	10	14	9	8	7	11	8
9	4	18	7	2	8	9	4	1
7	13	3	6	1	13	5	6	1
37	10	4	5	4	6	5	4	2
4	4	5	5	3	5	8	7	4
34	4	5	6	4	8	4	7	7
18	3	4	1	3	2	5	0	0
1	3	2	0	0	5	7	1	1
2	4	4	1	1	3	5	3	0
3	4	0	0	0	5	0	1	0
5	3	1	0	1	1	0	0	1
6	5	5	0	0	6	8	1	0
8	5	0	1	0	5	2	0	0
10	1	6	2	2	5	4	3	0
11	2	1	0	0	4	1	0	0
12	3	1	1	0	2	1	1	0
13	3	3	0	0	4	1	1	0
15	11	3	0	1	9	5	2	0
16	2	3	1	2	2	3	2	2
20	3	0	0	0	4	5	2	0
21	2	4	3	0	7	3	2	0
22	5	1	0	1	6	3	0	2
24	3	0	0	0	2	0	0	0
25	6	0	1	0	4	0	0	0
26	4	0	0	0	3	0	0	0
27	0	1	0	0	4	2	0	0
28	6	0	0	0	2	1	0	0
30	1	0	0	0	1	1	0	0
31	2	0	1	0	3	1	0	0
33	3	1	0	0	5	2	0	2
35	3	0	1	0	5	1	0	0
36	0	2	1	0	1	1	0	0
39	3	1	0	0	0	0	0	1
40	2	0	0	0	3	1	0	0
41	2	8	4	0	5	5	2	1
42	4	0	0	0	2	0	0	0
43	1	3	0	0	3	0	0	0
44	3	0	0	0	2	0	1	0
45	2	0	1	0	4	0	0	0
46	2	3	0	0	6	6	0	0
47	1	0	0	0	1	1	0	0
48	7	6	8	1	5	3	7	3
49	2	0	0	0	3	0	1	0
50	1	2	0	0	1	1	0	0
51	1	0	0	0	2	0	0	0
53	3	3	0	0	5	3	0	0
54	1	1	0	0	2	1	0	0
55	3	4	2	0	2	4	2	0
56	2	0	1	0	2	4	5	3
57	2	3	2	0	2	4	1	0
59	5	1	0	0	3	1	0	0
60	1	0	0	0	1	2	2	1

Tableau 12b. Nombre de fois qu'un item est identifié comme ayant un fonctionnement différentiel dans les échantillons de même taille avec le centile C90 ou les centiles C5 et C95 comme critère de décision (N = 100 échantillons).

Items	Méthode du delta de Mantel-Haenszel				Méthode de standardisation modifiée			
	250 sujets	500 sujets	1000 sujets	2000 sujets	250 sujets	500 sujets	1000 sujets	2000 sujets
19	41	67	86	98	38	71	88	99
17	27	33	51	65	30	48	62	68
14	31	38	52	77	24	36	49	70
23	12	13	16	24	26	32	49	66
58	34	53	71	78	16	17	23	26
52	13	24	26	32	21	30	33	37
38	19	25	28	35	12	22	22	30
32	8	21	17	16	10	23	19	16
29	18	21	41	31	12	13	22	17
9	12	28	15	9	20	30	26	22
7	20	17	19	12	27	21	22	28
37	16	7	10	8	22	15	15	14
4	10	20	22	20	13	14	15	22
34	8	8	14	13	12	9	11	8
18	7	7	4	5	10	8	8	17
1	11	6	5	1	9	4	4	1
2	13	19	8	5	11	15	7	7
3	8	2	0	0	10	3	1	2
5	8	3	1	3	12	11	5	4
6	13	19	14	8	11	14	8	6
8	4	2	2	0	14	4	6	4
10	3	11	4	3	9	17	20	8
11	4	1	0	0	5	6	4	0
12	3	4	1	0	2	6	1	0
13	5	4	2	0	12	6	8	4
15	14	7	2	2	19	12	11	5
16	6	7	13	8	12	16	18	11
20	5	0	0	0	8	1	2	2
21	5	5	5	0	8	9	13	2
22	7	5	0	3	8	7	5	6
24	5	0	0	0	5	2	0	0
25	8	0	1	0	12	2	5	1
26	3	2	0	0	9	4	4	0
27	3	1	0	0	9	8	0	0
28	13	5	1	0	8	4	1	0
30	3	1	0	0	6	2	0	1
31	7	1	2	0	3	2	1	0
33	4	2	0	0	16	7	0	0
35	4	0	2	0	7	2	2	0
36	1	3	1	0	8	7	2	1
39	5	5	0	0	5	4	0	3
40	3	0	1	0	8	2	1	0
41	9	12	9	8	3	7	3	1
42	5	0	0	0	9	1	0	0
43	6	3	0	1	7	2	0	0
44	7	1	1	0	6	2	1	0
45	7	2	3	0	4	1	1	0
46	12	9	3	3	3	4	0	1
47	2	2	0	0	4	4	0	0
48	12	11	12	10	11	8	7	4
49	7	2	1	0	4	2	1	0
50	1	3	2	0	9	6	2	0
51	9	11	7	4	2	0	1	0
53	8	4	4	0	0	2	1	0
54	8	4	3	0	2	1	0	0
55	8	10	5	0	1	5	1	0
56	12	9	6	5	1	0	1	0
57	20	16	18	24	0	0	0	0
59	16	7	2	1	2	1	0	0
60	14	6	0	1	0	0	0	0

Tableau 12b. (suite) Nombre de fois qu'un item est identifié comme ayant un fonctionnement différentiel dans les échantillons de même taille avec le centile C90 ou les centiles C5 et C95 comme critère de décision (N = 100 échantillons).

Items	Méthode du khi carré de Mantel-Haenszel				Méthode de régression logistique			
	250 sujets	500 sujets	1000 sujets	2000 sujets	250 sujets	500 sujets	1000 sujets	2000 sujets
19	43	71	87	98	31	67	87	95
17	33	41	66	76	36	34	68	67
14	28	39	48	71	25	40	41	67
23	17	23	28	40	15	18	26	36
58	25	40	56	68	30	44	57	75
52	20	25	33	33	18	25	34	30
38	22	33	28	38	14	21	22	30
32	11	27	25	26	12	22	28	28
29	21	23	33	26	14	21	28	28
9	16	26	20	14	16	21	18	8
7	22	14	17	9	14	13	13	6
37	18	12	14	17	14	7	11	10
4	8	15	13	15	10	17	17	19
34	11	10	17	14	16	10	17	16
18	8	7	6	7	5	9	7	4
1	5	3	2	1	14	9	5	3
2	10	8	6	4	8	9	6	5
3	6	0	0	0	10	5	2	0
5	10	3	3	3	8	4	2	1
6	10	18	8	3	14	13	7	5
8	9	2	2	0	8	4	0	0
10	4	13	9	3	9	9	6	4
11	5	3	0	0	5	3	1	0
12	5	5	2	0	4	8	1	0
13	7	6	3	2	8	6	5	3
15	17	9	4	2	15	8	10	5
16	8	8	13	5	5	9	9	5
20	8	1	0	0	10	9	4	4
21	6	9	6	1	10	6	6	1
22	9	5	3	3	8	7	0	4
24	5	0	0	0	4	1	0	0
25	9	0	1	0	10	0	2	0
26	6	4	0	0	4	1	0	0
27	3	3	0	0	7	3	0	0
28	15	5	0	0	8	2	0	0
30	4	1	0	0	3	4	0	0
31	5	1	2	0	5	2	0	0
33	7	3	0	0	12	4	1	2
35	6	1	2	0	9	1	2	0
36	3	3	1	0	4	3	0	0
39	5	6	1	0	0	1	0	1
40	5	0	0	0	6	1	0	0
41	6	11	8	4	10	12	6	4
42	7	1	0	0	9	2	0	0
43	4	4	0	1	5	3	0	1
44	10	2	1	0	4	2	1	0
45	7	2	4	0	7	1	2	0
46	4	6	1	2	9	8	1	2
47	3	1	0	0	5	3	1	0
48	11	12	11	7	11	9	11	5
49	5	2	1	0	6	1	1	0
50	4	3	1	0	5	3	1	0
51	4	4	2	1	3	4	2	1
53	8	4	2	0	8	11	4	2
54	5	2	0	0	5	2	0	0
55	7	8	2	0	9	10	8	8
56	5	2	3	0	6	10	13	9
57	3	8	5	5	7	9	1	3
59	10	3	1	1	8	4	1	0
60	5	0	0	0	6	5	4	3

Tableau 12c. Nombre de fois qu'un item est identifié comme ayant un fonctionnement différentiel dans les échantillons de même taille avec le centile C80 ou les centiles C10 et C90 comme critère de décision (N = 100 échantillons).

Items	Méthode du delta de Mantel-Haenszel				Méthode de standardisation modifiée			
	250 sujets	500 sujets	1000 sujets	2000 sujets	250 sujets	500 sujets	1000 sujets	2000 sujets
19	55	82	93	100	63	85	94	100
17	41	50	75	85	45	60	81	84
14	46	61	76	95	42	55	72	94
23	28	37	46	58	46	56	66	85
58	48	73	86	92	26	42	57	59
52	30	39	46	64	42	45	58	71
38	36	54	57	70	28	46	57	58
32	19	30	34	37	28	41	35	41
29	34	43	62	67	24	31	47	55
9	26	40	41	43	33	47	51	60
7	33	35	31	42	38	45	43	55
37	26	22	28	31	34	29	30	44
4	26	33	46	55	23	30	45	48
34	25	30	35	34	25	16	28	26
18	18	12	14	21	21	24	24	44
1	20	16	11	5	18	15	6	4
2	21	25	21	24	21	25	20	21
3	17	8	2	1	21	7	5	5
5	15	14	7	7	31	24	21	14
6	27	36	45	31	22	33	34	33
8	11	7	4	4	22	12	8	13
10	10	21	20	11	22	31	39	37
11	10	7	3	1	23	12	10	4
12	12	14	9	4	12	20	7	6
13	14	12	10	5	22	22	13	10
15	25	19	15	10	28	29	26	24
16	19	31	34	28	22	27	31	26
20	12	1	0	0	17	9	11	4
21	11	13	16	7	14	22	29	14
22	17	16	16	14	17	22	22	21
24	12	7	1	1	13	6	4	3
25	14	6	4	1	20	14	9	3
26	9	9	4	1	19	15	9	2
27	10	6	0	0	18	15	2	1
28	19	11	5	0	19	7	5	5
30	13	3	2	2	13	6	2	1
31	14	5	6	1	13	6	4	1
33	14	4	1	0	30	15	7	5
35	10	3	5	0	15	5	8	1
36	9	8	4	0	16	10	5	3
39	15	11	7	6	13	17	14	7
40	12	2	2	0	19	11	5	1
41	21	24	24	34	6	15	13	6
42	14	3	0	0	20	4	0	2
43	16	9	0	2	19	8	4	1
44	14	4	1	0	12	7	2	1
45	12	9	6	0	10	3	3	0
46	19	20	16	9	9	9	4	2
47	8	5	2	0	9	10	1	0
48	20	20	28	21	19	21	22	13
49	19	10	6	0	14	6	4	1
50	8	6	2	1	18	11	5	2
51	18	18	13	11	4	3	3	1
53	15	11	10	6	5	9	2	0
54	20	9	7	3	7	2	0	0
55	15	19	11	4	5	10	5	0
56	23	20	18	15	6	7	3	2
57	27	32	34	51	2	4	1	0
59	30	17	7	4	8	2	0	1
60	25	19	8	2	4	1	0	0

Tableau 12c. (suite) Nombre de fois qu'un item est identifié comme ayant un fonctionnement différentiel dans les échantillons de même taille avec le centile C80 ou les centiles C10 et C90 comme critère de décision (N = 100 échantillons).

Items	Méthode du khi carré de Mantel-Haenszel				Méthode de régression logistique			
	250 sujets	500 sujets	1000 sujets	2000 sujets	250 sujets	500 sujets	1000 sujets	2000 sujets
19	58	80	93	100	49	81	92	98
17	48	56	83	94	46	51	80	91
14	42	53	73	94	43	55	70	88
23	34	44	57	66	27	37	49	64
58	43	63	83	91	48	64	85	94
52	32	46	54	69	26	42	47	63
38	33	52	58	72	35	48	50	59
32	25	38	42	52	25	40	49	50
29	34	40	59	58	28	39	55	55
9	27	40	43	41	26	35	36	35
7	34	30	29	40	28	30	31	35
37	29	29	36	43	32	19	31	29
4	23	30	38	42	21	31	36	44
34	25	35	42	45	22	26	40	42
18	19	16	14	25	17	17	12	21
1	17	11	4	3	22	17	9	6
2	18	23	17	18	22	24	18	18
3	11	2	1	0	27	11	6	5
5	17	14	7	6	18	12	5	5
6	23	30	33	22	22	27	25	15
8	16	9	4	5	22	7	7	3
10	11	23	25	13	17	24	16	11
11	12	10	8	2	10	9	7	3
12	14	17	10	5	10	17	5	2
13	21	17	13	6	15	14	15	9
15	28	20	19	11	26	17	23	21
16	19	29	33	24	16	20	26	24
20	15	5	1	0	20	17	17	14
21	17	18	22	10	16	15	15	8
22	20	17	18	18	19	16	12	12
24	14	5	1	1	12	6	1	0
25	15	5	5	2	18	6	6	2
26	10	12	5	1	11	6	3	0
27	12	10	1	1	14	9	3	2
28	19	14	6	1	17	13	3	1
30	12	4	2	1	12	5	1	1
31	12	5	6	2	13	6	6	1
33	19	6	2	0	26	15	9	13
35	13	8	6	0	14	5	4	0
36	16	9	5	0	12	11	5	0
39	16	17	8	7	10	8	4	4
40	17	2	2	0	12	5	3	0
41	16	24	24	25	15	22	17	15
42	15	4	0	0	21	6	3	1
43	20	11	1	2	10	6	1	1
44	15	7	2	1	13	6	4	1
45	12	10	6	0	10	9	6	2
46	14	16	10	5	21	17	6	5
47	8	10	1	1	13	11	7	1
48	25	25	28	24	24	21	26	26
49	16	9	7	0	11	6	7	0
50	12	8	2	1	10	8	4	0
51	10	15	9	8	10	15	11	3
53	14	9	8	5	9	20	14	7
54	13	7	2	1	12	4	2	1
55	12	15	8	5	20	22	18	23
56	14	11	7	7	19	25	32	38
57	14	15	15	21	20	17	12	11
59	21	9	2	3	18	10	4	4
60	11	4	1	0	18	18	10	13

Tableau 12d. Nombre de fois qu'un item est identifié comme ayant un fonctionnement différentiel dans les échantillons de même taille avec des critères fixes, soit la valeur critique au seuil de 0,05, l'unité ou 0,075 selon les méthodes (N = 100 échantillons).

Items	Méthode du delta de Mantel-Haenszel				Méthode de standardisation modifiée			
	250 sujets	500 sujets	1000 sujets	2000 sujets	250 sujets	500 sujets	1000 sujets	2000 sujets
19	66	77	87	95	72	78	88	94
17	51	45	56	57	54	57	67	64
14	57	54	63	70	52	54	50	59
23	40	33	22	20	55	50	55	58
58	61	71	75	75	41	38	35	24
52	44	39	32	22	55	42	38	36
38	44	48	34	34	37	40	36	22
32	27	27	17	10	34	37	23	15
29	43	42	44	30	32	30	25	12
9	39	33	20	6	43	46	32	24
7	44	31	21	10	54	39	31	27
37	41	19	13	6	41	28	18	11
4	43	34	25	21	36	27	22	16
34	39	26	20	12	33	15	15	4
18	24	11	8	5	30	24	10	10
1	25	16	6	1	25	16	4	0
2	30	26	11	7	28	18	12	7
3	27	4	1	0	27	7	1	0
5	25	8	3	3	40	25	9	6
6	37	34	19	9	30	31	17	5
8	24	8	3	0	34	14	6	3
10	19	21	9	1	33	27	23	7
11	15	8	0	0	29	13	6	1
12	23	13	3	0	21	16	1	0
13	23	10	3	0	34	17	9	4
15	30	17	5	1	35	26	14	6
16	33	21	13	5	38	27	17	6
20	15	1	0	0	22	9	3	1
21	19	15	6	0	23	21	14	2
22	28	17	1	2	23	18	8	0
24	20	2	0	0	22	4	1	0
25	19	3	0	0	27	13	4	0
26	11	7	0	0	21	9	3	0
27	16	5	0	0	27	12	1	1
28	24	11	1	0	25	9	2	0
30	22	3	1	0	21	6	1	0
31	21	4	2	0	21	7	1	0
33	26	4	0	0	36	14	0	0
35	22	3	1	0	21	8	2	0
36	23	5	1	0	25	10	2	0
39	21	12	1	0	20	14	5	0
40	25	2	0	0	29	7	2	0
41	31	25	13	3	14	12	4	1
42	23	1	0	0	27	4	0	0
43	25	8	0	0	24	8	0	0
44	19	3	1	0	19	5	1	0
45	18	8	4	0	14	4	1	0
46	30	20	5	3	14	7	0	1
47	17	7	1	0	19	11	0	0
48	29	18	13	4	24	20	12	3
49	29	10	1	0	21	7	0	0
50	14	4	1	0	25	8	1	0
51	30	21	10	2	10	1	0	0
53	23	10	3	0	9	5	1	0
54	29	10	2	0	11	1	0	0
55	21	18	2	0	16	9	1	0
56	35	17	10	2	15	2	1	0
57	40	33	25	21	8	1	0	0
59	40	15	2	1	15	1	0	0
60	39	18	1	1	8	1	0	0

Tableau 12d. (suite) Nombre de fois qu'un item est identifié comme ayant un fonctionnement différentiel dans les échantillons de même taille avec des critères fixes, soit la valeur critique au seuil de 0,05, l'unité ou 0,075 selon les méthodes (N = 100 échantillons).

Items	Méthode du khi carré de Mantel-Haenszel				Méthode de régression logistique			
	250 sujets	500 sujets	1000 sujets	2000 sujets	250 sujets	500 sujets	1000 sujets	2000 sujets
19	33	75	94	100	36	77	95	100
17	31	46	86	98	35	44	83	97
14	27	44	76	100	27	50	72	99
23	15	33	58	90	15	32	56	86
58	24	51	86	99	34	60	90	100
52	21	32	57	92	21	36	61	89
38	19	37	62	95	16	39	57	90
32	13	30	48	86	15	34	58	89
29	13	31	66	88	17	31	64	88
9	12	31	45	82	15	30	42	73
7	18	20	38	76	17	23	38	71
37	15	18	47	76	17	14	39	71
4	6	17	45	74	14	25	46	75
34	8	16	47	74	17	19	53	77
18	7	11	24	60	11	14	20	57
1	6	8	8	11	13	15	16	23
2	5	17	23	47	11	22	32	53
3	5	1	1	4	10	8	13	27
5	4	7	13	24	6	10	10	18
6	7	22	40	55	13	21	36	49
8	3	6	6	19	8	5	8	16
10	2	17	27	46	9	19	27	41
11	6	8	11	15	8	8	12	13
12	2	10	11	20	6	13	9	18
13	9	11	15	32	9	8	19	36
15	15	15	25	41	15	16	28	61
16	5	14	36	60	6	17	37	58
20	5	3	4	4	11	15	25	41
21	5	15	26	34	12	11	26	35
22	9	11	23	50	12	15	22	42
24	4	1	1	4	4	3	2	4
25	5	2	5	7	11	3	9	6
26	5	7	7	2	5	3	5	4
27	3	4	2	2	6	8	6	11
28	9	8	9	18	7	6	5	15
30	2	3	6	7	3	5	2	6
31	4	2	7	12	5	5	11	11
33	7	4	8	8	15	11	14	40
35	4	3	7	8	7	4	7	6
36	5	5	2	4	2	9	6	4
39	2	10	12	19	1	6	9	18
40	2	1	3	6	7	3	6	5
41	6	18	27	54	11	20	27	47
42	6	0	1	4	9	4	7	8
43	5	5	4	13	7	4	2	16
44	7	3	7	6	6	2	9	7
45	5	4	7	12	4	7	9	17
46	3	9	17	27	10	15	17	20
47	3	5	2	4	9	8	8	12
48	8	13	28	52	12	18	35	58
49	4	6	7	8	6	4	8	3
50	3	4	3	9	3	7	6	5
51	4	10	11	22	5	12	15	24
53	5	4	13	16	7	15	22	35
54	2	3	2	5	4	4	5	7
55	4	9	13	21	9	17	28	58
56	1	6	12	23	8	20	42	69
57	5	8	21	56	10	13	22	42
59	7	4	3	9	9	8	11	12
60	2	2	0	2	5	15	19	33

Un examen rapide des quatre tableaux révèle une relation entre le nombre de fois qu'un item est détecté et la grandeur de l'échantillon, et ce, quelle que soit la méthode d'analyse du fonctionnement différentiel des items. Il révèle aussi une relation entre la grandeur des différences de fonctionnement estimées pour la population et la fréquence de détection dans les 100 échantillons de même taille. En général, plus la différence de fonctionnement estimée pour la population est importante, plus la fréquence de détection des items est grande. Un examen plus attentif des tableaux montre également une relation entre le critère de décision et la fréquence de détection des items. Dans le cas des centiles, les tendances observées se ressemblent pour toutes les méthodes d'analyse du fonctionnement différentiel considérées. Dans le cas des critères fixés a priori, les tendances observées diffèrent selon que la méthode repose sur une mesure de l'ampleur des différences dans le fonctionnement des items (la méthode du delta de Mantel-Haenszel et la méthode de standardisation modifiée) ou sur un khi carré (la méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique).

Dans le cas de critères qui reposent sur des valeurs à un point centile prédéterminé, nous constatons que la fréquence de détection des items qui ont un fonctionnement différentiel assez élevé dans la population (les items 19, 17, 14, 23 et 58) tend à croître avec l'augmentation de la grandeur des échantillons, peu importe le centile utilisé comme critère de décision. Cette tendance s'observe également pour les items avec un fonctionnement différentiel faible (les items 52, 38, 32, 29 et 9) ou très faible (les items 7, 37, 4, 34 et 18) lorsqu'on se sert des centiles C_{10} et C_{90} ou C_{80}. Dans le cas des centiles $C_{2,5}$ et $C_{97,5}$ ou C_{95} et des centiles C_5 et C_{95} ou C_{90}, la fréquence de détection des items avec un fonctionnement différentiel faible ou très faible tend à se maintenir au même niveau ou à épouser une forme curviligne au lieu de croître de façon continue avec l'accroissement de la taille des échantillons. Un effet de plafonnement est possible

eu égard au fait que les critères retenus ne permettent de détecter que cinq ou 10 pour cent d'items et que le test compte 25 pour cent d'items avec un fonctionnement différentiel non négligeable dans la population. Quel que soit le centile utilisé comme critère de décision, la fréquence de détection des items qui ont un fonctionnement différentiel négligeable dans la population tend à décroître avec l'accroissement du nombre de sujets dans les échantillons. Bien sûr, il s'agit là de tendances générales, et certains items ne suivent pas la tendance. Certaines divergences sont liées à des particularités de la méthode d'analyse du fonctionnement différentiel, par exemple, la détection de l'item 57 avec la méthode du delta de Mantel-Haenszel. Toutefois, les items qui ne suivent pas la tendance générale sont, plus souvent qu'autrement, ceux qui affichent les différences de fonctionnement les plus grandes parmi les items identifiés comme ayant un fonctionnement différentiel négligeable dans la population.

En ce qui concerne les critères fixés a priori, nous constatons que les méthodes qui font appel à un khi carré (la méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique) voient leurs fréquences de détection croître de façon systématique avec l'augmentation de la taille des échantillons. Il en est ainsi tant parmi les items qui ont un fonctionnement différentiel non négligeable dans la population que parmi ceux qui ont un fonctionnement différentiel négligeable. Dans le cas des méthodes qui utilisent une mesure de l'ampleur des différences dans le fonctionnement des items (la méthode du delta de Mantel-Haenszel et la méthode de standardisation modifiée), nous constatons un phénomène semblable à celui observé pour les centiles. Il y a augmentation constante de la fréquence de détection des items dans le cas des items qui affichent des différences de fonctionnement assez élevées dans la population. Par contre, les fréquences de détection des items tendent à décroître pour les items qui présentent des différences de fonctionnement négligeables, faibles ou très faibles.

À partir des différences de détection observées pour chaque item, il appert que la constance des décisions, d'un échantillon à un autre de même taille, sera d'autant plus grande que les items affichent un fonctionnement différentiel important dans la population. Corollaire inévitable, il faudra des échantillons d'autant plus grands que les différences de fonctionnement que l'on espère détecter dans la population seront plus petites. Avec une différence de fonctionnement de l'ordre de 0,10, l'item 19 est celui qui accuse la différence de fonctionnement la plus grande dans le test analysé. C'est aussi l'item qui offre les décisions les plus susceptibles d'être constantes. Des échantillons de 1000 sujets (500 personnes par groupe) fournissent des décisions d'une stabilité élevée. Les autres items exigent des échantillons beaucoup plus nombreux pour atteindre une stabilité comparable. Nous en déduisons que les données de la présente étude sur la fidélité et l'efficacité des méthodes sont étroitement liées à la grandeur des différences de fonctionnement des items qui accusent un fonctionnement différentiel non négligeable dans la population.

Il appert également que des échantillons de 250 sujets (125 personnes par groupe) ne peuvent donner lieu à des décisions fiables, lorsque les différences de fonctionnement que l'on espère pouvoir détecter dans la population sont peu élevées. Enfin, quelle que soit la taille des échantillons, il appert que l'on ne peut s'attendre à détecter les mêmes items d'un échantillon à un autre de même taille, lorsque les items affichent des différences de fonctionnement inférieures à 0,10 dans la population. Autrement dit, il est possible que les items détectés aient tous des différences de fonctionnement non négligeables dans la population, mais que ce ne soit pas les mêmes items qui soient détectés. Plus concrètement, il est possible que le nombre d'items détectés dans deux échantillons de même taille soit identique et que, parmi les items détectés, ceux qui se classent parmi les 15 items retenus comme ayant un fonctionnement différentiel

non négligeable dans la population ne soient pas les mêmes d'un échantillon à l'autre, même s'ils appartiennent tous à la catégorie des items avec un fonctionnement différentiel non négligeable.

Coefficients de validité des décisions

Afin d'avoir une idée de la validité des méthodes en ce qui concerne la justesse des décisions prises à partir d'échantillons de taille réduite, nous avons examiné la relation qui existe entre les items classés comme ayant un fonctionnement différentiel dans les échantillons et les 15 items qui accusent des différences de fonctionnement non négligeable dans la population. Pour ce faire, nous avons eu recours aux coefficients de corrélation phi. Les tableaux 13a, 13b, 13c et 13d résument les résultats des analyses corrélationnelles effectuées. Nous y indiquons l'étendue de la distribution et la médiane. Parce que celle-ci est peu représentative de l'ensemble des coefficients, nous avons également indiqué la valeur des coefficients au premier et au troisième quartile. Les tableaux 13a, 13b et 13c décrivent les coefficients de validité obtenus pour les critères de décision qui utilisent des centiles. Le tableau 13d dépeint les coefficients de validité pour les critères fixés a priori. La distribution des coefficients en fonction de leur grandeur est fournie à l'annexe B (voir les tableaux 8a, 8b, 8c, et 8d).

De façon générale, il appert que les critères décisionnels basés sur des centiles fournissent des coefficients de validité d'autant plus élevés que le centile permet d'identifier comme ayant un fonctionnement différentiel un nombre important d'items. Cette tendance s'explique par le fait qu'aucun des centiles considérés ne permet de déceler la totalité des items retenus comme ayant un fonctionnement différentiel non négligeable dans la population. Les centiles considérés permettent d'identifier cinq, 10 ou 20 pour cent des items qui composent le test. Or, 25 pour cent

des items se sont avérés avoir un fonctionnement différentiel non négligeable dans la population. Les critères utilisés engendrent forcément une certaine quantité de faux négatifs.

Nous avons estimé la valeur maximale des coefficients de validité lorsque le critère de décision est un centile. Pour ce faire, nous avons postulé qu'il n'y a pas de divergence entre les échantillons et la population pour ce qui est des items sans fonctionnement différentiel. Les centiles C_5 ou $C_{2,5}$ et $C_{97,5}$ permettent de détecter cinq pour cent des items qui composent le test. Dans ces conditions, on peut détecter trois items avec les méthodes qui font appel à un khi carré et quatre items avec les méthodes qui utilisent des indices orientés. Dans le premier cas, la valeur maximale possible du coefficient de validité est de 0,397 ; dans le second, elle s'élève à 0,463. Avec les centiles C_{10} ou C_5 et C_{95}, on peut identifier six items et la valeur maximale possible du coefficient de validité est de 0,577. Celle-ci s'élève à 0,866 pour les centiles C_{80} ou C_{10} et C_{90}, alors que l'on peut identifier 12 items. On comprendra que la valeur maximale possible ne peut être estimée facilement lorsqu'on se sert des critères fixés a priori.

En comparant les coefficients de validité obtenus lorsqu'on se sert des centiles comme critère de décision, il appert que l'on atteint la valeur maximale possible d'autant plus souvent et avec des grandeurs d'échantillon d'autant plus petites que le centile utilisé est moins englobant. On détecte moins d'items, mais une plus grande proportion d'items qui se classent parmi les items qui présentent un fonctionnement différentiel non négligeable dans la population.

À l'examen des coefficients de validité obtenus pour chaque méthode et chaque grandeur d'échantillon, il appert également que la grandeur des coefficients de validité tend à croître avec la taille des échantillons. Non seulement la grandeur des coefficients tend à croître, mais aussi le

nombre de coefficients significatifs au seuil de 0,05. Parallèlement à cette augmentation, leur variabilité tend à décroître. Il en est ainsi quelle que soit la méthode et le critère de décision considérés. Ceci dit, il y a beaucoup de chevauchement entre les distributions de coefficients de validité pour les 100 échantillons de même taille et, somme toute, peu de différence entre les méthodes pour un même critère de décision.

En se basant sur l'erreur-type du coefficient de corrélation phi et la médiane de chaque distribution de coefficients, on note des différences significatives en fonction de la taille des échantillons. Quels que soient la méthode d'analyse du fonctionnement différentiel utilisée et le critère de décision, il n'y a pas de différence significative entre les coefficients médians de deux grandeurs d'échantillon adjacentes. Par contre, il y en a lorsqu'on oppose les valeurs médianes des coefficients de validité des échantillons de 250 sujets à celles des échantillons de 1000 et de 2000 sujets ou lorsqu'on oppose les valeurs médianes des coefficients des échantillons de 500 sujets à celles des échantillons de 2000 sujets. Les différences ne sont toutefois pas systématiques.

Lorsqu'on utilise les centiles C_{90} ou C_5 et C_{95} , la différence entre les coefficients médians est significative peu importe la méthode pour les échantillons de 250 sujets opposés aux échantillons de 1000 ou de 2000 sujets. Il en va de même pour les centiles C_{80} ou C_{10} et C_{90}. Pour les critères fixés a priori, seule la méthode de standardisation modifiée (DDSM) fait exception. Dans ce cas, il n'y a pas de différence significative entre le coefficient médian des échantillons de 250 sujets et celui des échantillons de 1000 sujets. Par contre, il y en a une lorsqu'on oppose le coefficient médian des échantillons de 250 sujets à celui des échantillons de 2000 sujets.

La situation diffère lorsqu'on oppose les échantillons de 500 sujets aux échantillons de 2000 sujets. Dans le cas des centiles C_{90} ou C_5 et C_{95} , seule la méthode du khi carré de Mantel-Haenszel présente des coefficients médians dont la différence est significative. Dans le cas des centiles C_{80} ou C_{10} et C_{90}, toutes les méthodes présentent des valeurs médianes significativement différentes, à l'exception de la méthode du khi carré de Mantel-Haenszel. Dans le cas des critères fixés a priori, seules la méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique affichent des différences significatives.

Dans le cas des centiles C_5 ou $C_{2,5}$ et $C_{97,5}$, seules quelques valeurs médianes accusent des différences significatives. Dans le cas du delta de Mantel-Haenszel, cela se produit lorsqu'on oppose la médiane des échantillons de 2000 sujets à celle des échantillons de 250 sujets. Il en va de même pour la méthode de standardisation modifiée et la méthode de régression logistique. Cette dernière affiche aussi une différence significative lorsqu'on oppose la valeur médiane des échantillons de 1000 sujets à celle des échantillons de 250 sujets. Dans tous les autres cas, il n'y a pas de différence significative.

Reste à déterminer quelle méthode est la plus susceptible d'offrir des décisions valides. L'examen de l'étendue des distributions, des médianes et des valeurs au premier et au troisième quartile donne à penser qu'il y a peu de différence entre les méthodes pour des échantillons de même taille. On retiendra seulement que la méthode de standardisation modifiée se comporte de façon analogue à la méthode du delta de Mantel-Haenszel. Par ailleurs, la méthode de régression logistique pourrait donner des décisions un peu moins valides que les autres méthodes.

Tableau 13a. Validité des décisions basées sur les centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$ par rapport aux items qui ont été classés comme ayant un fonctionnement différentiel non négligeable dans la population (N = 100 comparaisons pour chaque grandeur d'échantillon).

Indices	Échantillons			
	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}				
Étendue	de -0,154 à 0,463*	de -0,035 à 0,522*	de 0,000 à 0,463*	de 0,154 à 0,522*
Médiane	0,154	0,309*	0,309*	0,463*
1er quartile	0.000	0,154	0,309*	0,309*
3e quartile	0,309*	0,309*	0,463*	0,463*
DDSM				
Étendue	de -0,154 à 0,463*	de -0,035 à 0,463*	de 0,000 à 0,463*	de 0,154 à 0,522*
Médiane	0,154	0,309*	0,309*	0,463*
1er quartile	0.000	0,154	0,309*	0,364*
3e quartile	0,309*	0,383*	0,463*	0,463*
X^2_{MH}				
Étendue	de -0,132 à 0,397*	de 0,044 à 0,397*	de 0,044 à 0,397*	de 0,221 à 0,397*
Médiane	0,221	0,221	0,397*	0,397*
1er quartile	0,044	0,221	0,221	0,397*
3e quartile	0,221	0,397*	0,397*	0,397*
$X^2_{amél.}$				
Étendue	de -0,132 à 0,397*	de -0,132 à 0,397*	de 0,044 à 0,397*	de 0,044 à 0,397*
Médiane	0,044	0,221	0,397*	0,397*
1er quartile	0,044	0,044	0,221	0,397*
3e quartile	0,221	0,397*	0,397*	0,397*

* Le coefficient de corrélation est significatif au seuil de 0,05.

** Les coefficients de validité sont des coefficients de corrélation de phi.

*** La valeur maximale possible des coefficients de validité est de 0,463 pour le Δ_{MH} et la DDSM et de 0,397 pour le X^2_{MH} et le $X^2_{amél.}$

Tableau 13b. Validité des décisions basées sur les centiles C_{90} ou C_5 et C_{95} par rapport aux items qui ont été classés comme ayant un fonctionnement différentiel non négligeable dans la population (N = 100 comparaisons pour chaque grandeur d'échantillon).

Indices	Échantillons			
	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}				
Étendue	de -0192 à 0,577*	de -0,064 à 0,577*	de 0,030 à 0,629*	de 0,192 à 0,679*
Médiane	0,192	0,321*	0,449*	0,449*
1er quartile	0,064	0,192	0,321*	0,434*
3e quartile	0,270*	0,321*	0,449*	0,577*
DDSM				
Étendue	de -0,192 à 0,577*	de -0,064 à 0,577*	de 0,030 à 0,629*	de 0,192 à 0,629*
Médiane	0,192	0,321*	0,449*	0,479*
1er quartile	0,064	0.192	0,321*	0,449*
3e quartile	0,321*	0,449*	0,449*	0,577*
X^2_{MH}				
Étendue	de -0,192 à 0,577*	de -0,064 à 0,577*	de 0,030 à 0,629*	de 0,192 à 0,577*
Médiane	0,192	0,321*	0,449*	0,577*
1er quartile	0,064	0,250	0,321*	0,449*
3e quartile	0,321*	0,449*	0,449*	0,577*
$X^2_{amél.}$				
Étendue	de -0,192 à 0,577*	de -0,064 à 0,577*	de 0,064 à 0,577*	de 0,192 à 0,577*
Médiane	0,192	0,321*	0,449*	0,449*
1er quartile	0,064	0,192	0,321*	0,449*
3e quartile	0,225	0,353*	0,449*	0,577*

* Le coefficient de corrélation est significatif au seuil de 0,05.

** Les coefficients de validité sont des coefficients de corrélation phi.

*** La valeur maximale possible des coefficients de validité est de 0,577 pour toutes les méthodes à l'étude.

Tableau 13c. Validité des décisions basées sur les centiles C_{80} ou C_{10} et C_{90} par rapport aux items qui ont été classés comme ayant un fonctionnement différentiel non négligeable dans la population (N = 100 comparaisons pour chaque grandeur d'échantillon).

Indices	Échantillons			
	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}				
Étendue	de -0,192 à 0,481*	de 0,000 à 0,577*	de 0,192 à 0,770*	de 0,257* à 0,867*
Médiane	0,192	0,289*	0,444*	0,577*
1er quartile	0,096	0,257*	0,385*	0,481*
3e quartile	0,289*	0,385*	0,577*	0,674*
DDSM				
Étendue	de -0,192 à 0,481*	de 0,070 à 0,674*	de 0,192 à 0,770*	de 0,257* à 0,867*
Médiane	0,192*	0,287*	0,481*	0,577*
1er quartile	0,096	0,257*	0,385*	0,481*
3e quartile	0,289*	0,385*	0,577*	0,674*
X^2_{MH}				
Étendue	de -0,192 à 0,577*	de 0,096 à 0,674*	de 0,192 à 0,866*	de 0,289 à 0,866*
Médiane	0,192	0,350*	0,481*	0,577*
1er quartile	0,096	0,289*	0,385*	0,481*
3e quartile	0,289*	0,385*	0,577*	0,577*
$X^2_{amél.}$				
Étendue	de -0,192* à 0,577*	de 0,000 à 0,577*	de 0,192* à 0,770*	de 0,385* à 0,867*
Médiane	0,192	0,289*	0,481*	0,577*
1er quartile	0,096	0,192	0,385*	0,481*
3e quartile	0,289*	0,385*	0,547*	0,674*

* Le coefficient de corrélation est significatif au seuil de 0,05.

** Les coefficients de validité sont des coefficients de corrélation phi.

*** La valeur maximale possible des coefficients de validité est de 0,866 pour toutes les méthodes à l'étude.

Tableau 13d. Validité des décisions basées sur des critères fixés a priori par rapport aux items qui ont été classés comme ayant un fonctionnement différentiel non négligeable dans la population (N = 100 comparaisons pour chaque grandeur d'échantillon).

Indices	Échantillons			
	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}				
Étendue	de -0,210 à 0,517*	de -0,117 à 0,629*	de 0,081 à 0,728*	de 0,192 à 0,728*
Médiane	0,174	0,293*	0,449*	0,453*
1er quartile	0,094	0,220	0,335*	0,383*
3e quartile	0,294*	0,404*	0,510*	0,522*
DDSM				
Étendue	de -0,210 à 0,537*	de 0,113 à 0,622*	de 0,052 à 0,775*	de -0.075 à 0,728*
Médiane	0,200	0,296*	0,423*	0,463*
1er quartile	0,122	0,226	0,340*	0,395*
3e quartile	0,322*	0,423*	0,510*	0,522*
X^2_{MH}				
Étendue	de -0,174 à 0,566*	de -0,064 à 0,728*	de 0,155 à 0,911*	de 0,314* à 0,833*
Médiane	0,221	0,321*	0,501*	0,599*
1er quartile	0,104	0,226	0,383*	0,525*
3e quartile	0,321*	0,404*	0,577*	0,661*
$X^2_{amél.}$				
Étendue	de -0,174 à 0,522*	de -0,090 à 0,671*	de 0,111 à 0,783*	de 0,251 à 0,765*
Médiane	0,154	0,270*	0,410*	0,523*
1er quartile	0,050	0,192	0,348*	0,447*
3e quartile	0,296*	0,404*	0,501*	0,597*

* Le coefficient de corrélation est significatif au seuil de 0,05.

** Les coefficients de validité sont des coefficients de corrélation phi.

Nombre de décisions correctes

Nous avons également déterminé le nombre de décisions correctes pour chaque échantillon, chaque méthode, chaque critère de décision et chaque grandeur d'échantillon. Ce nombre correspond au nombre d'items détectés dans l'échantillon qui accusent des différences de fonctionnement non négligeables dans la population et au nombre d'items non détectés qui n'ont pas été retenus comme ayant un fonctionnement différentiel dans la population. Plus succinctement, il s'agit de la somme des vrais positifs et des vrais négatifs. Les tableaux 14a, 14b, 14c et 14d font état de la moyenne et de l'écart-type des sommes ainsi calculées pour chaque méthode, chaque grandeur d'échantillon et chaque critère de décision. Nous y indiquons aussi la valeur minimale et la valeur maximale du nombre de décisions correctes ainsi que la médiane.

Les données relatives au nombre de décisions correctes confirment les tendances observées à partir des coefficients de validité. De façon générale, le nombre de décisions correctes tend à croître avec l'augmentation du nombre de sujets dans les échantillons. Les moyennes compilées pour chaque grandeur d'échantillon en témoignent. Par ailleurs, la variabilité du nombre de décisions correctes pour chaque échantillon de même taille tend à décroître avec l'accroissement de la taille des échantillons. Ceci dit, il y a beaucoup de chevauchement dans les distributions de fréquences associées à chaque grandeur d'échantillon. Il en est ainsi pour toutes les méthodes à l'étude et tous les critères de décision considérés.

À partir des données recueillies, un échantillon de 250 sujets peut donner un nombre de décisions correctes aussi grand qu'un échantillon de 2000 sujets. Toutefois, la possibilité d'atteindre un

nombre aussi élevé de décisions correctes est beaucoup moins grande. Par ailleurs, des échantillons de 2000 sujets ne donneront pas un nombre de décisions correctes aussi faible que les fréquences les plus faibles atteintes avec des échantillons de 250 sujets.

Même si le nombre de décisions correctes tend à augmenter avec l'accroissement de la taille des échantillons, l'augmentation demeure somme toute assez limitée. Considérons la situation lorsqu'on passe d'échantillons de 250 sujets à des échantillons de 2000 sujets. Lorsqu'on se sert des centiles C_5 ou $C_{2,5}$ et $C_{97,5}$, l'augmentation de la fréquence moyenne varie entre 4,100 et 6,017 pour cent selon les méthodes. Dans le cas des centiles C_{10} ou C_5 et C_{95}, elle se situe entre 7,750 et 8,472 pour cent. Pour les centiles C_{80} ou C_{10} et C_{90}, l'augmentation se situe entre 11,200 et 15,300 pour cent selon les méthodes. Autrement dit, plus le centile utilisé comme critère de décision permet d'identifier une proportion importante d'items, plus l'augmentation du nombre de sujets peut accroître le nombre de décisions correctes.

Dans le cas des critères fixés a priori, l'augmentation de la fréquence moyenne de décisions correctes est de 13,050 pour cent avec la méthode du delta de Mantel-Haenszel et de 14,316 pour cent avec la méthode de standardisation modifiée. Par contre, l'augmentation n'est que de 2,834 pour cent pour la méthode de régression logistique et de 5,767 pour cent pour la méthode du khi carré de Mantel-Haenszel. En somme, l'augmentation de la taille des échantillons profite davantage aux méthodes qui utilisent une mesure de l'ampleur des différences dans le fonctionnement des items. Il en est ainsi parce que ces méthodes permettent un nombre de décisions correctes généralement moins grand que la méthode du khi carré de Mantel-Haenszel ou la méthode de régression logistique lorsque les échantillons ne comptent que 250 sujets.

Autre fait à noter, il y a peu de différence entre les méthodes d'analyse du fonctionnement différentiel des items à l'étude pour ce qui est du nombre de décisions correctes lorsqu'on utilise un centile comme critère de décision. De fait, la différence entre la fréquence moyenne de la méthode qui offre le plus de décisions correctes et celle qui en offre le moins varie de 0,530 à, au plus, 3,417 pour cent, selon la taille des échantillons et les centiles utilisés. La faiblesse de l'écart montre que les méthodes sont, à toutes fins utiles, aussi valides l'une que l'autre pour ce qui est du nombre de décisions correctes.

Dans le cas des critères fixés a priori, la méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique sont en mesure de fournir plus de décisions valides avec des échantillons de 250 sujets. La différence entre les fréquences moyennes pour les 100 échantillons de même taille varie entre 5,200 et 8,233 pour cent. Pour les échantillons de 500 sujets ou plus, il y a très peu de différences entre les méthodes à l'étude, peu importe la méthode considérée. Tout au plus peut-on noter que la méthode de régression logistique tend à offrir moins de décisions correctes au fur et à mesure que la taille des échantillons augmente, un peu comme si elle était plus lente à réagir à l'augmentation de la taille des échantillons.

La présente étude avait pour but de comparer l'efficacité de la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991a, 1991b, 1992a, 1992b, 1993a et 1995) à l'efficacité de la méthode de Mantel-Haenszel et de la méthode de régression logistique. Sous ce rapport, il appert que la méthode de standardisation modifiée fournisse des décisions dont la validité s'apparente généralement à celle des deltas de Mantel-Haenszel. Dans ces conditions, on peut s'attendre à des taux de détection assez semblables.

Tableau 14a. Fréquence de décisions correctes basées sur les centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$ par rapport aux items classés comme ayant un fonctionnement différentiel non négligeable dans la population (N = 100 comparaisons pour chaque grandeur d'échantillon).

Indices	Échantillons			
	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}				
Moyenne	**44,98**	**46,34**	**47,59**	**48,29**
Écart-type	1,853	1,782	1,564	1,113
Étendue	de 41,00 à 49,00	de 42,00 à 50,00	de 43,00 à 49,00	de 45,00 à 50,00
Médiane	45,00	47,00	47,00	49,00
DDSM				
Moyenne	**44,88**	**46,53**	**47,55**	**48,49**
Écart-type	1,986	1,920	1,424	1,049
Étendue	de 41,00 à 49,00	de 42,00 à 49,00	de 43,00 à 49,00	de 45,00 à 50,00
Médiane	45,00	47,00	47,00	49,00
X^2_{MH}				
Moyenne	**45,36**	**46,54**	**47,40**	**47,82**
Écart-type	1,795	1,445	0,995	0,575
Étendue	de 42,00 à 48,00	de 44,00 à 48,00	de 44,00 à 48,00	de 46,00 à 48,00
Médiane	46,00	46,00	48,00	48,00
$X^2_{amél.}$				
Moyenne	**44,96**	**46,14**	**47,20**	**47,66**
Écart-type	1,740	1,614	1,206	0,807
Étendue	de 42,00 à 48,00	de 42,00 à 48,00	de 44,00 à 48,00	de 44,00 à 48,00
Médiane	44,00	46,00	48,00	48,00

NB : Compte tenu des centiles utilisés comme critère de décision, le nombre maximal de décisions correctes est de 49 (81,167 %) avec le Δ_{MH} ou la DDSM et de 48 (80 %) avec le X^2_{MH} ou $X^2_{amél.}$ pour la régression logistique. Ce nombre peut être dépassé à l'occasion, lorsque deux items se situent au même centile.

Tableau 14b. Fréquence de décisions correctes basées sur les centiles C_{90} ou C_5 et C_{95} par rapport aux items classés comme ayant un fonctionnement différentiel non négligeable dans la population (N = 100 comparaisons pour chaque grandeur d'échantillon).

Indices	Échantillons			
	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}				
Moyenne	**44,45**	**46,55**	**48,31**	**49,33**
Écart-type	2,380	2,157	1,921	1,658
Étendue	de 39,00 à 51,00	de 41,00 à 51,00	de 42,00 à 52,00	de 45,00 à 53,00
Médiane	45,00	47,00	49,00	49,00
DDSM				
Moyenne	**44,79**	**46,66**	**48,16**	**49,66**
Écart-type	2,426	2,147	2,024	1,506
Étendue	de 39,00 à 51,00	de 41,00 à 51,00	de 42,00 à 52,00	de 45,00 à 52,00
Médiane	45,00	47,00	49,00	49,50
X^2_{MH}				
Moyenne	**45,03**	**47,11**	**48,31**	**50,04**
Écart-type	2,595	2,128	1,921	1,348
Étendue	de 39,00 à 51,00	de 41,00 à 51,00	de 42,00 à 52,00	de 45,00 à 51,00
Médiane	45,00	47,00	49,00	51,00
$X^2_{amél.}$				
Moyenne	**44,39**	**46,38**	**48,48**	**49,38**
Écart-type	2,386	2,286	1,812	1,600
Étendue	de 39,00 à 51,00	de 41,00 à 51,00	de 43,00 à 51,00	de 45,00 à 51,00
Médiane	45,00	47,00	49,00	49,00

NB : Compte tenu des centiles utilisés comme critère de décision, le nombre maximal de décisions correctes est de 51(85 %), quelle que soit la méthode. Ce nombre peut être dépassé à l'occasion, lorsque deux items se situent au même centile.

Tableau 14c. Fréquence de décisions correctes basées sur les centiles C_{80} ou C_{10} et C_{90} par rapport aux items classés comme ayant un fonctionnement différentiel non négligeable dans la population (N = 100 comparaisons pour chaque grandeur d'échantillon).

Indices	Échantillons			
	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}				
Moyenne	**42,75**	**45,71**	**48,23**	**50,67**
Écart-type	3,006	2,571	2.767	2,598
Étendue	de 35,00 à 49,00	de 39,00 à 51,00	de 43,00 à 55,00	de 44,00 à 57,00
Médiane	43,00	45,00	48,00	51,00
DDSM				
Moyenne	**44,51**	**45,83**	**48,60**	**51,23**
Écart-type	2,397	2,674	2,617	2,490
Étendue	de 38,00 à 50,00	de 40,00 à 53,00	de 43,00 à 55,00	de 44,00 à 57,00
Médiane	45,00	45,00	49,00	51,00
X^2_{MH}				
Moyenne	**43,10**	**46.01**	**49,60**	**50,36**
Écart-type	3,249	2,691	3,023	2,476
Étendue	de 35,00 à 51,00	de 41,00 à 53,00	de 43,00 à 57,00	de 45,00 à 57,00
Médiane	43,00	46,00	49,00	51,00
$X^2_{amél.}$				
Moyenne	**42,46**	**45,30**	**48,25**	**51,64**
Écart-type	2,938	2,702	2,844	2,376
Étendue	de 35,00 à 51,00	de 39,00 à 51,00	de 43,00 à 55,00	de 47,00 à 57,00
Médiane	43,00	45,00	49,00	51,00

NB : Compte tenu des centiles utilisés comme critère de décision, le nombre maximal de décisions correctes est de 57 (95 %), quelle que soit la méthode. Ce nombre peut être dépassé à l'occasion, lorsque deux items se situent au même centile.

Tableau 14d. Fréquence de décisions correctes basées sur des critères fixés a priori par rapport aux items classés comme ayant un fonctionnement différentiel non négligeable dans la population (N = 100 comparaisons pour chaque grandeur d'échantillon).

Indices		Échantillons			
		250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}					
	Moyenne	**40,48**	**45,64**	**48,53**	**49,07**
	Écart-type	3,735	2,904	2,303	1,760
	Étendue	de 31,00 à 48,00	de 36,00 à 52,00	de 42,00 à 54,00	de 45,00 à 54,00
	Médiane	41,00	45,00	49.00	49,00
DDSM					
	Moyenne	**41,39**	**46,00**	**48,55**	**49,22**
	Écart-type	3,536	2,554	2,226	1,751
	Étendue	de 32,00 à 50,00	de 39,00 à 52,00	de 41,00 à 55,00	de 44,00 à 54,00
	Médiane	42,00	46,00	48,00	49,00
X^2_{MH}					
	Moyenne	**45,42**	**46,64**	**48,63**	**48,88**
	Écart-type	2,142	2,646	3,113	2,735
	Étendue	de 40,00 à 51,00	de 39,00 à 54,00	de 42,00 à 58,00	de 41,00 à 56,00
	Médiane	45,00	46,00	49,00	49,00
$X^2_{amél.}$					
	Moyenne	**44,51**	**45,66**	**46,80**	**46,28**
	Écart-type	2,397	2,606	3,111	3,042
	Étendue	de 38,00 à 50,00	de 40,00 à 53,00	de 39,00 à 55,00	de 38,00 à 54,00
	Médiane	45,00	45,50	46,00	46,00

Efficacité relative des méthodes

Un premier aperçu de l'efficacité relative de la méthode de Mantel-Haenszel (Δ_{MH} et X^2_{MH}), de la méthode de standardisation modifiée (DDSM) et de la méthode de régression logistique ($X^2_{amél.}$) nous a été fourni par l'étude de la validité et de la fidélité des décisions. Toutefois, pour avoir une idée plus juste de l'efficacité des méthodes, nous avons déterminé, pour chaque méthode, chaque critère décisionnel et chaque grandeur d'échantillon, le taux de détection des items qui ont un fonctionnement différentiel non négligeable dans la population et le taux de détection des items faussement identifiés comme ayant un fonctionnement différentiel.

Les taux de détection ont été étudiés sous deux angles distincts : le taux relatif par échantillon et le taux objectif par échantillon. Dans les pages qui suivent, nous préciserons la notion de taux relatif et nous ferons état des taux observés pour chaque méthode, chaque critère de décision et chaque grandeur d'échantillon. Par la suite, nous définirons la notion de taux objectif et nous étudierons les résultats obtenus sous ce rapport.

Les taux relatifs de détection

Les taux relatifs de détection consistent en des taux établis en fonction de la totalité des items détectés dans les échantillons. Ils présupposent une répartition des données dans un tableau à deux cases et comportent deux taux possibles : un taux d'items détectés qui ont un fonctionnement différentiel non négligeable dans la population (items avec FD) et un taux d'items détectés qui ont un fonctionnement différentiel négligeable dans cette même population

(items sans FD). Pour simplifier la discussion, nous parlerons d'identification correcte (IC) dans le premier cas et d'identification incorrecte (INC) dans le second.

	Items dans la population		
	avec FD	sans FD	
Items détectés dans les échantillons	Identification correcte (IC)	Identification incorrecte (INC)	Total des items détectés
	15 items	45 items	

Le taux d'identification correcte correspond à la proportion moyenne d'items détectés qui ont un fonctionnement différentiel dans la population, par rapport à la totalité des items détectés dans les 100 échantillons de même grandeur.

$$Taux_{IC} = \sum [\frac{nombre_{IC}}{total_{détecté}}] \ / \ 100 \qquad (26)$$

Le taux d'identification incorrecte correspond à la proportion moyenne d'items détectés qui n'ont pas de fonctionnement différentiel dans la population, par rapport à la totalité des items détectés dans les 100 échantillons de même grandeur.

$$Taux_{INC} = \sum [\frac{nombre_{INC}}{Total_{détecté}}] \ / \ 100 \qquad (27)$$

Les tableaux 15a, 15b, 15c et 15d reproduisent les taux d'identification incorrecte et les taux d'identification correcte. Les premiers correspondent aux items avec un fonctionnement

différentiel négligeable ; les seconds, aux items avec un fonctionnement différentiel non négligeable. Pour ces derniers, nous indiquons également la répartition des taux de détection selon que les items ont un fonctionnement différentiel assez élevé, faible ou très faible. Le tableau 7 (p. 81) précise les caractéristiques de ces trois groupes d'items.

Les figures 2 à 4 font état des taux d'identification correcte et des taux d'identification incorrecte par rapport à des centiles prédéterminés. La figure 5 fait état des taux d'identification correcte ou incorrecte par rapport à des critères fixés a priori. Nous examinerons d'abord les taux basés sur la valeur des indices à un centile prédéterminé, puis les taux basés sur des critères fixés a priori. Les taux d'identification incorrecte correspondent à la différence entre les taux d'identification correcte et l'unité. Pour cette raison, nous ne nous y attarderons pas.

Tableau 15a. Taux moyens d'identification correcte et d'identification incorrecte par rapport à la totalité des items détectés dans un échantillon en appliquant les centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$ comme critère de décision (N = 100 échantillons).

Indice	FD	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}	**Négligeable**	0,50250	0,33498	0,17706	0,09383
	Non négligeable	**0,49750**	**0,66503**	**0,82294**	**0,90617**
	Très faible	0,10000	0,07882	0,08479	0,06667
	Faible	0,13250	0,19458	0,19701	0,17284
	Assez élevé	0,26500	0,39163	0,54115	0,66667
DDSM	**Négligeable**	0,51574	0,31204	0,18362	0,07125
	Non négligeable	**0,48426**	**0,68796**	**0,81638**	**0,92875**
	Très faible	0,13317	0,09582	0,12655	0,10811
	Faible	0,11380	0,20639	0,18610	0,17199
	Assez élevé	0,23729	0,38575	0,50372	0,64865
X^2_{MH}	**Négligeable**	0,44000	0,24333	0,10265	0,03000
	Non négligeable	**0,56000**	**0,75667**	**0,89735**	**0,97000**
	Très faible	0,11333	0,07000	0,07616	0,05000
	Faible	0,14667	0,24333	0,20199	0,15333
	Assez élevé	0,30000	0,44333	0,61921	0,76667
$X^2_{amél.}$	**Négligeable**	0,50667	0,31000	0,13333	0,05667
	Non négligeable	**0,49333**	**0,69000**	**0,86667**	**0,94333**
	Très faible	0,11333	0,09000	0,08000	0,04667
	Faible	0,13000	0,19333	0,18667	0,18000
	Assez élevé	0,25000	0,40667	0,60000	0,71667

NB : Les taux en caractère gras correspondent aux taux d'identification correcte. Sous ces taux, on indique la répartition en fonction de l'importance des différences dans le fonctionnement des items. Au dessus, apparaissent les taux d'identification incorrecte.

Tableau 15b. Taux moyens d'identification correcte et d'identification incorrecte par rapport à la totalité des items détectés dans un échantillon en appliquant les centiles C_{90} ou C_5 et C_{95} comme critère de décision (N = 100 échantillons).

Indice	FD	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}	**Négligeable**	0,54531	0,37274	0,23002	0,14682
	Non négligeable	**0,45470**	**0,62726**	**0,76998**	**0,85318**
	Très faible	0,10049	0,09688	0,11256	0,09462
	Faible	0,11532	0,19540	0,20718	0,20065
	Assez élevé	0,23888	0,33498	0,45025	0,55791
DDSM	**Négligeable**	0,51730	0,36438	0,24183	0,12052
	Non négligeable	**0,48270**	**0,63562**	**0,75817**	**0,87948**
	Très faible	0,13839	0,10948	0,11601	0,14495
	Faible	0,12356	0,19281	0,19935	0,19870
	Assez élevé	0,22076	0,33333	0,44281	0,53583
X^2_{MH}	**Négligeable**	0,49751	0,32446	0,18303	0,08000
	Non négligeable	**0,50249**	**0,67554**	**0,81697**	**0,92000**
	Très faible	0,11111	0,09651	0,11148	0,10333
	Faible	0,14925	0,22296	0,23128	0,22833
	Assez élevé	0,24212	0,35607	0,47421	0,58833
$X^2_{amél.}$	**Négligeable**	0,55075	0,38500	0,21000	0,13500
	Non négligeable	**0,44925**	**0,61500**	**0,79000**	**0,86500**
	Très faible	0,09870	0,09333	0,10833	0,09167
	Faible	0,12313	0,18333	0,21667	0,20667
	Assez élevé	0,22795	0,33833	0,46500	0,56667

NB : Les taux en caractère gras correspondent aux taux d'identification correcte. Sous ces taux, on indique la répartition en fonction de l'importance des différences dans le fonctionnement des items. Au dessus, apparaissent les taux d'identification incorrecte.

Tableau 15c. Taux moyens d'identification correcte et d'identification incorrecte par rapport à la totalité des items détectés dans un échantillon en appliquant les centiles C_{80} ou C_{10} et C_{90} comme critère de décision (N = 100 échantillons).

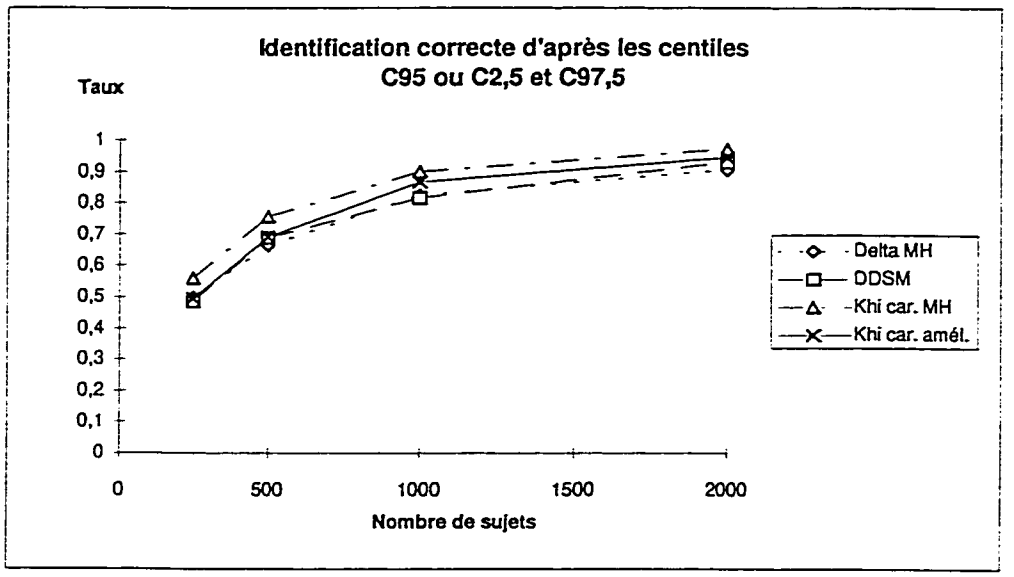
Indice	FD	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}	**Négligeable**	0,59321	0,47069	0,36730	0,26781
	Non négligeable	**0,40679**	**0,52932**	**0,63270**	**0,73219**
	Très faible	0,10605	0,10900	0,12654	0,14988
	Faible	0,12013	0,17011	0,19721	0,23014
	Assez élevé	0,18061	0,25021	0,30896	0,35217
DDSM	**Négligeable**	0,57367	0,46601	0,35197	0,24571
	Non négligeable	**0,42634**	**0,53399**	**0,64803**	**0,75429**
	Très faible	0,11605	0,11794	0,13980	0,17714
	Faible	0,12757	0,17199	0,20395	0,23265
	Assez élevé	0,18272	0,24406	0,30428	0,34449
X^2_{MH}	**Négligeable**	0,57904	0,45802	0,33056	0,22333
	Non négligeable	**0,42097**	**0,54198**	**0,66944**	**0,77667**
	Très faible	0,10815	0,11638	0,13239	0,16250
	Faible	0,12562	0,17955	0,21316	0,24333
	Assez élevé	0,18719	0,24605	0,32390	0,37083
$X^2_{amél.}$	**Négligeable**	0,60583	0,48750	0,36470	0,27667
	Non négligeable	**0,39417**	**0,51250**	**0,63530**	**0,72333**
	Très faible	0,10000	0,10250	0,12490	0,14250
	Faible	0,11667	0,17000	0,19734	0,21833
	Assez élevé	0,17750	0,24000	0,31307	0,36250

NB : Les taux en caractère gras correspondent aux taux d'identification correcte. Sous ces taux, on indique la répartition en fonction de l'importance des différences dans le fonctionnement des items. Au-dessus, apparaissent les taux d'identification incorrecte.

Tableau 15d. Taux moyens d'identification correcte et d'identification incorrecte par rapport à la totalité des items détectés dans un échantillon en appliquant des critères fixés a priori comme critère de décision (N = 100 échantillons).

Indice	FD	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}	**Négligeable**	0,62711	0,47133	0,25520	0,12245
	Non négligeable	**0,37289**	**0,52867**	**0,74480**	**0,87755**
	Très faible	0,10742	0,10842	0,12067	0,10019
	Faible	0,11080	0,16936	0,20388	0,18924
	Assez élevé	0,15467	0,25090	0,42025	0,58813
DDSM	**Négligeable**	0,60624	0,45496	0,25850	0,10189
	Non négligeable	**0,39376**	**0,54505**	**0,74150**	**0,89811**
	Très faible	0,11419	0,11982	0,13061	0,12830
	Faible	0,11831	0,17568	0,20952	0,20566
	Assez élevé	0,16127	0,24955	0,40136	0,56415
X^2_{MH}	**Négligeable**	0,45643	0,40000	0,36989	0,41150
	Non négligeable	**0,54357**	**0,60000**	**0,63011**	**0,58850**
	Très faible	0,11203	0,10000	0,14409	0,16423
	Faible	0,16183	0,19634	0,19928	0,20210
	Assez élevé	0,26971	0,30366	0,28674	0,22217
$X^2_{amél.}$	**Négligeable**	0,53695	0,46667	0,44260	0,47329
	Non négligeable	**0,46305**	**0,53333**	**0,55740**	**0,52671**
	Très faible	0,11463	0,09596	0,12500	0,14649
	Faible	0,12670	0,17172	0,17985	0,17905
	Assez élevé	0,22172	0,26566	0,25255	0,20117

NB : Les taux en caractère gras correspondent aux taux d'identification correcte. Sous ces taux, on indique la répartition en fonction de l'importance des différences dans le fonctionnement des items. Au-dessus, apparaissent les taux d'identification incorrecte.



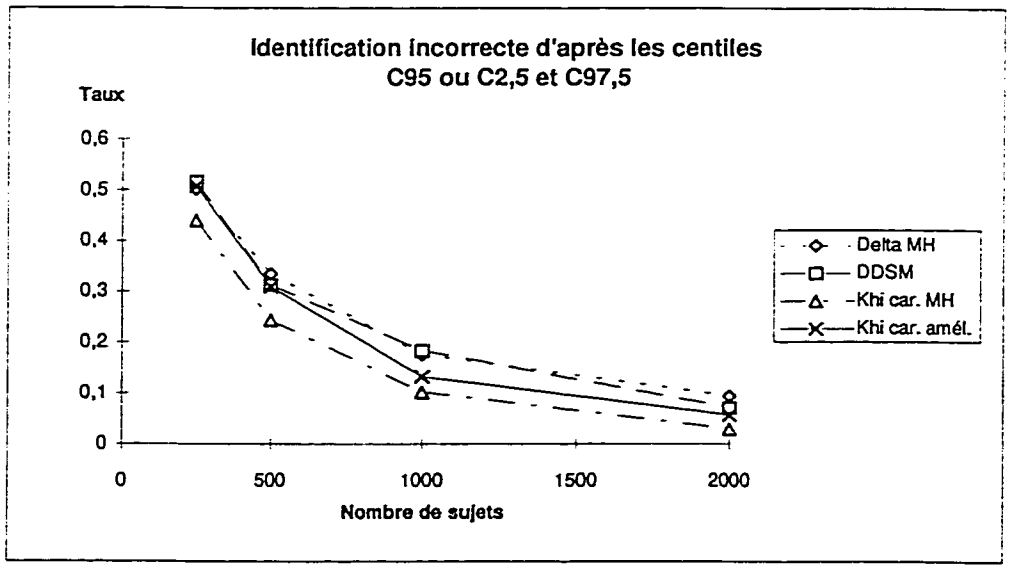
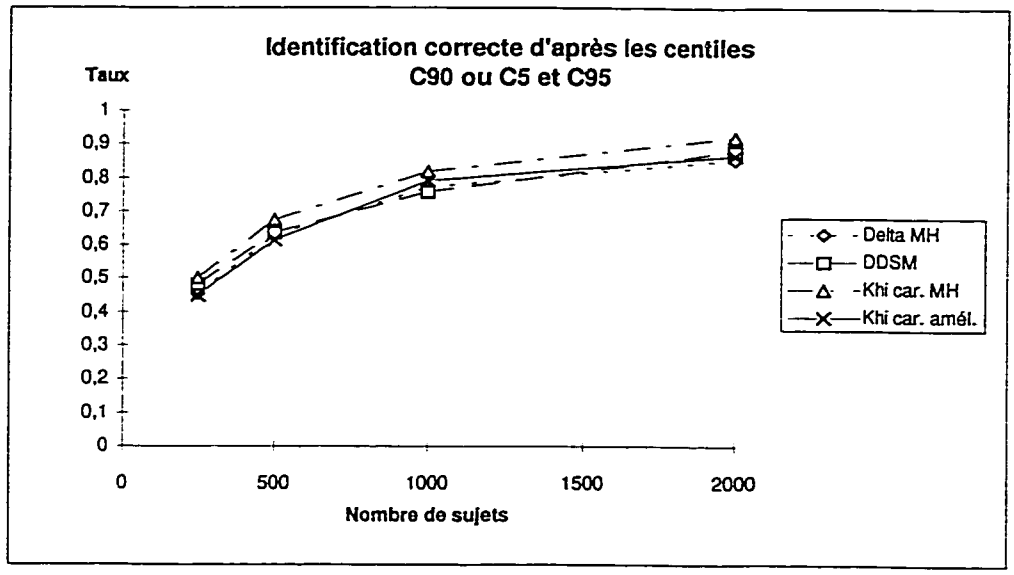


Figure 2. Taux moyens observés d'identification correcte ou incorrecte pour les critères C95 ou C2,5 et C97,5.



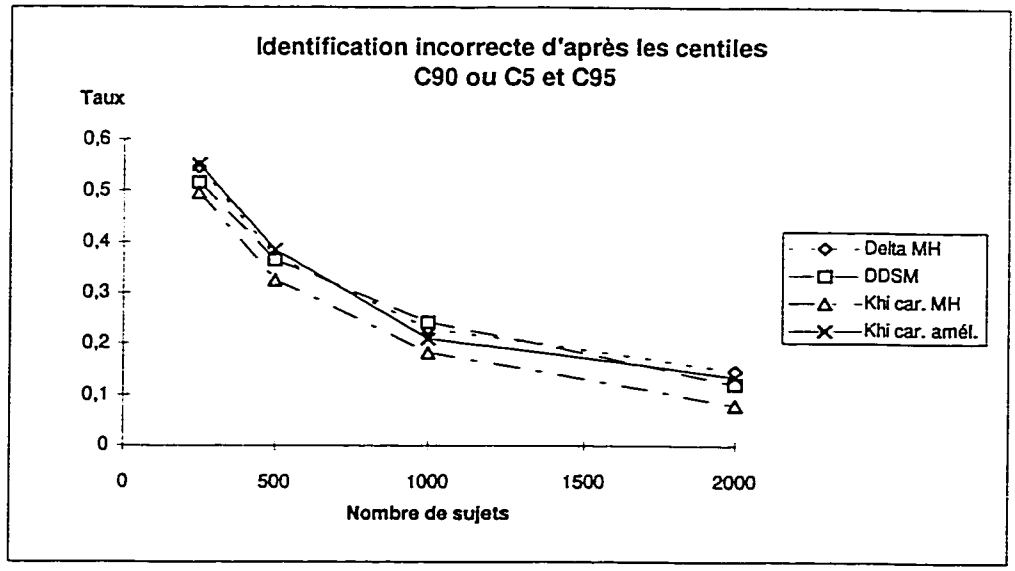
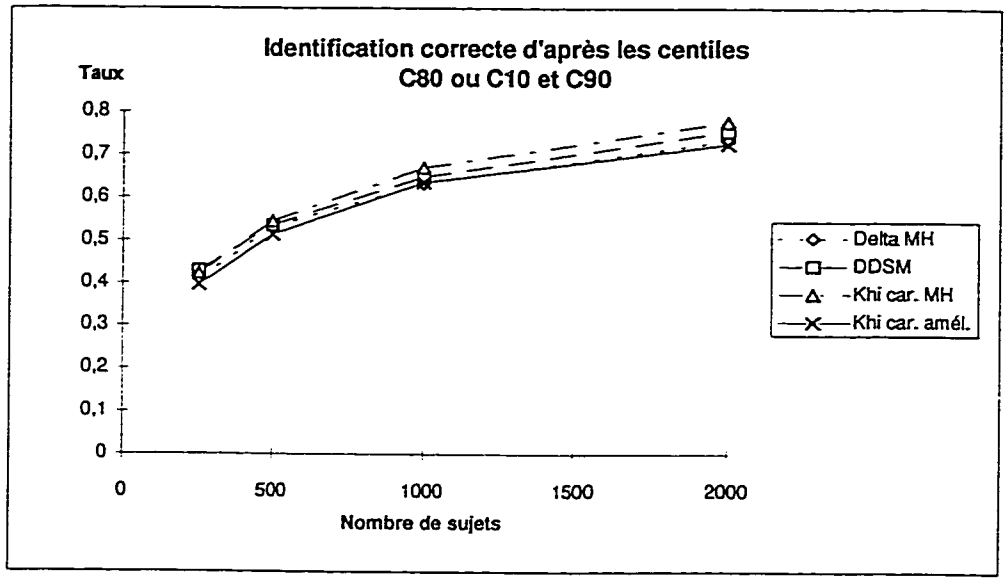


Figure 3. Taux moyens observés d'identification correcte ou incorrecte pour les critères C90 ou C5 et C95.



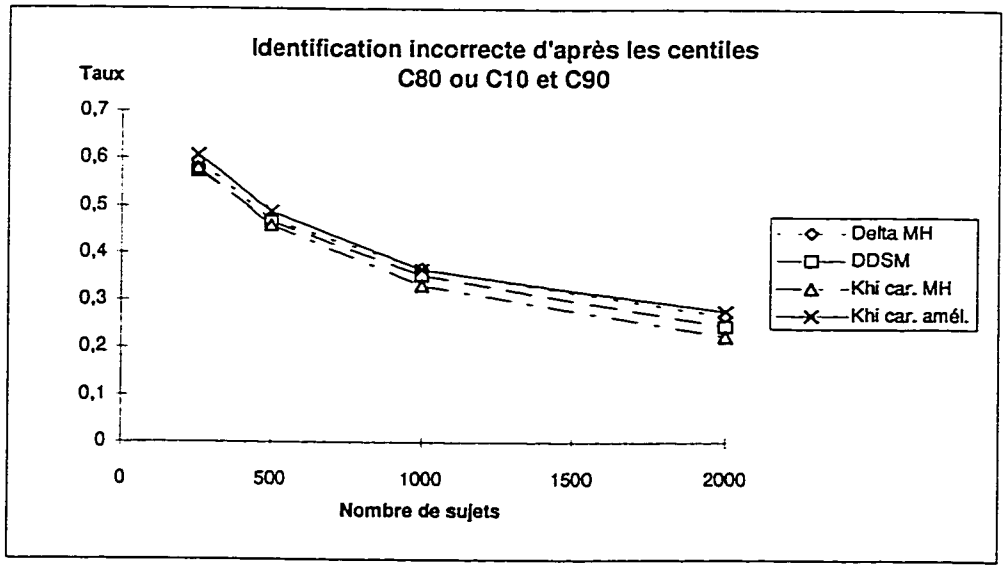
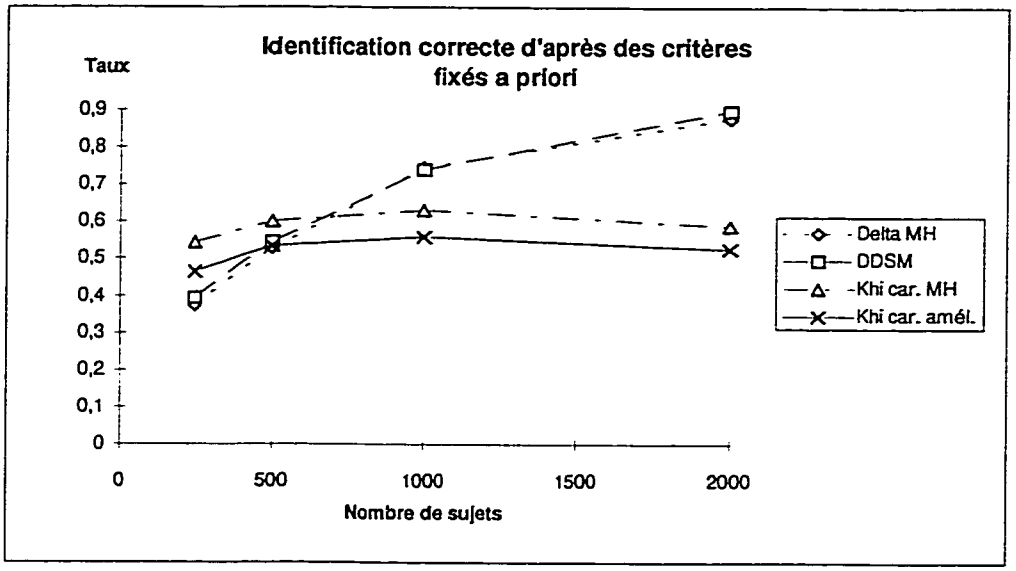


Figure 4. Taux moyens observés d'identification correcte ou incorrecte pour les critères C80 ou C10 et C90.



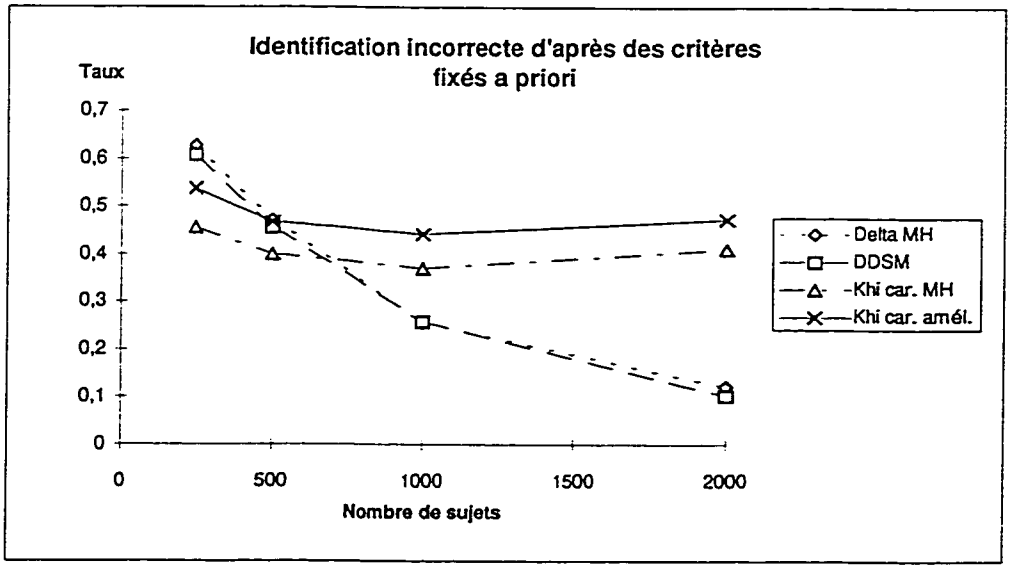


Figure 5. Taux moyens observés d'identification correcte ou incorrecte pour des critères fixés a priori.

Taux relatifs de détection basés sur des centiles prédéterminés

La figure 2 fait état des taux moyens d'identification correcte lorsque le critère utilisé pour classer les items est le centile C_{95} pour les indices non orientés (X^2_{MH} et $X^2_{amél.}$) et les centiles $C_{2,5}$ ou $C_{97,5}$ pour les indices orientés (Δ_{MH} et DDSM). Dans ces conditions, il n'y a que cinq pour cent des items qui peuvent être identifiés comme ayant un fonctionnement différentiel. Le test analysé compte 60 items. Règle générale, quatre items sont classés comme ayant un fonctionnement différentiel dans le cas des indices orientés et trois dans le cas des indices non orientés.

La figure 3 reproduit les taux moyens d'identification correcte lorsque le critère décisionnel est le centile C_{90} pour les indices non orientés et les centiles C_5 ou C_{95} pour les indices orientés. Dans ces conditions, 10 pour cent des items peuvent être identifiés comme ayant un fonctionnement différentiel. Le test comptant 60 items, six items sont ainsi détectés, parfois sept lorsque deux items se situent au même centile. Le même nombre d'items est détecté, qu'il s'agisse d'indices orientés ou non orientés. Toutefois, il arrive plus souvent que l'on détecte sept items au lieu de six dans le cas des indices orientés.

La figure 4 rend compte des taux moyens d'identification correcte lorsque le critère de décision utilisé est le centile C_{80} pour les indices non orientés et les centiles C_{10} ou C_{90} pour les indices orientés. Avec ce critère, 20 pour cent des items peuvent être identifiés comme ayant un fonctionnement différentiel. Pour le test analysé, 12 items sont détectés dans chaque échantillon, peu importe le type d'indices. Comme précédemment, il arrive que l'on détecte 13 items au lieu

de 12, et cela est plus fréquent dans le cas des indices orientés. Quels que soient les indices, la détection de 13 items au lieu de 12 demeure cependant l'exception.

Prises globalement, les trois figures montrent que le taux d'identification correcte varie non seulement en fonction de la taille des échantillons, mais aussi en fonction du critère de décision. Règle générale, celui-ci augmente avec l'utilisation d'un centile plus extrême (ici, C_{95} ou $C_{2,5}$ et $C_{97,5}$). Moins d'items sont détectés. Parmi les items détectés, le risque d'erreur est moins grand et la probabilité d'identifier des items qui ont un fonctionnement différentiel non négligeable dans la population croît en conséquence. Les coefficients de validité des décisions reflétaient le phénomène.

Non seulement le taux d'identification correcte augmente avec l'utilisation d'un critère plus extrême, mais l'écart entre les méthodes tend à s'accroître. La différence est plus accentuée avec l'utilisation des centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$ et moins accentuée avec l'utilisation des centiles C_{80} ou C_{10} et C_{90}. Toutefois, peu importe le critère de décision, nous observons la même tendance : une augmentation très importante du taux de détection lorsqu'on passe d'échantillons de 250 sujets à des échantillons de 500 sujets, une augmentation un peu moins importante lorsqu'on passe d'échantillons de 500 sujets à des échantillons de 1000 sujets et une augmentation nettement moins accentuée lorsqu'on passe d'échantillons de 1000 à 2000 sujets.

Enfin, le classement des méthodes en fonction de leur efficacité dépend de la grandeur des échantillons et du critère utilisé. La méthode du khi carré de Mantel-Haenszel est généralement la plus efficace. Celle-ci possède les taux les plus élevés pour tous les critères décisionnels et

toutes les grandeurs d'échantillons, mis à part les échantillons de 250 sujets lorsqu'on utilise le centile C_{80} comme critère de décision. C'est au niveau des autres méthodes que se manifestent les différences d'efficacité.

L'utilisation des centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$ donne lieu à peu de différence entre les méthodes pour les échantillons de 250 sujets. La méthode de régression logistique occupe le deuxième rang pour les échantillons de 500, de 1000 et de 2000 sujets. La méthode de standardisation modifiée occupe le troisième rang pour les échantillons de 500 et de 2000 sujets et le quatrième pour les échantillons de 1000 sujets. La méthode du delta de Mantel-Haenszel occupe le troisième rang pour les échantillons de 1000 sujets et le quatrième pour les échantillons de 500 et de 2000 sujets.

Avec l'utilisation des centiles C_{80} ou C_{10} et C_{90} comme critère de décision, la méthode de standardisation modifiée occupe le second rang pour les échantillons de 500, de 1000 et de 2000 sujets. Elle se classe au premier rang pour les échantillons de 250 sujets. La méthode du delta de Mantel-Haenszel vient au troisième rang pour toutes les grandeurs d'échantillon, sauf pour les échantillons de 1000 sujets où elle passe au quatrième rang. La méthode de régression logistique occupe le quatrième rang pour toutes les grandeurs d'échantillon, mis à part les échantillons de 1000 sujets où elle vient au troisième rang.

La situation la plus ambiguë est celle qui se produit avec l'utilisation des centiles C_{90} ou C_5 et C_{95} comme critère décisionnel. Pour les échantillons de 250, de 500 et de 2000 sujets, la méthode de standardisation modifiée vient au deuxième rang. Elle passe au quatrième rang pour les

échantillons de 1000 sujets. La méthode du delta de Mantel-Haenszel vient au troisième rang pour toutes les grandeurs d'échantillon à l'exception des échantillons de 2000 sujets, où elle se classe au dernier rang. La méthode de régression logistique vient au quatrième rang pour les échantillons de 250 et de 500 sujets. Par la suite, elle passe au deuxième rang pour les échantillons de 1000 sujets et au quatrième rang pour les échantillons de 2000 sujets.

Règle générale, l'écart entre les taux d'identification correcte est faible pour une même grandeur d'échantillon. Lorsque nous utilisons les centiles C_{90} ou C_5 et C_{95} (figure 3), il varie de six à huit pour cent entre la méthode la plus efficace et la méthode la moins efficace, selon la grandeur des échantillons. La méthode du khi carré de Mantel-Haenszel exige des échantillons de 500 sujets pour que le taux d'identification correcte s'élève à 68 pour cent. Les autres méthodes y parviennent avec des échantillons de 675 sujets. Pour que le taux d'identification correcte s'élève à 80 pour cent, la méthode du khi carré de Mantel-Haenszel exige des échantillons de 925 sujets ; la méthode de standardisation modifiée demande des échantillons de 1100 sujets ; la méthode du delta de Mantel-Haenszel et la méthode de régression logistique, des échantillons d'environ 1300 sujets.

Lorsque nous utilisons les centiles C_{80} ou C_{10} et C_{90} comme critère de décision, le taux d'identification correcte ne peut atteindre le seuil de 80 pour cent, et ce, quelle que soit la méthode. Pour parvenir à un taux de 68 pour cent, la méthode du khi carré de Mantel-Haenszel nécessite des échantillons de 1100 sujets. La méthode de standardisation modifiée exige des échantillons de 1300 sujets. La méthode du delta de Mantel-Haenszel et la méthode de régression logistique requièrent des échantillons d'environ 1500 sujets.

Nous nous sommes montrés très exigeants pour identifier les items qui ont un fonctionnement différentiel dans la population, puisque nous avons considéré comme ayant un fonctionnement différentiel tout item dont la différence de difficulté standardisée (DDS) est égale ou supérieure à 0,05 en valeur absolue, en arrondissant au centième le plus proche. Parmi les items qui répondent à ce critère, certains accusent une différence assez élevée ; d'autres, une différence faible ; et d'autres, une différence très faible. On peut se demander ce qui arriverait si nous ne considérions comme ayant un fonctionnement différentiel non négligeable que les items dont la différence de fonctionnement est assez élevée. On peut également s'interroger sur la répartition des taux de détection en fonction de la grandeur des différences dans le fonctionnement des items.

L'examen des tableaux 15a, 15b et 15c (p. 135 à 137) montre que le taux de détection des items qui ont un fonctionnement différentiel assez élevé est nettement plus grand que le taux de détection des items qui ont un fonctionnement différentiel faible ou très faible. En fait, le taux de détection des items qui ont un fonctionnement différentiel assez élevé représente plus de la moitié du taux observé pour l'ensemble des items avec un fonctionnement différentiel non négligeable. Si nous ne considérions comme ayant un fonctionnement différentiel non négligeable dans la population que les items avec une différence de fonctionnement assez élevée, le taux d'identification correcte serait donc nettement inférieur et le taux d'identification incorrecte se verrait augmenter des taux de détection observés pour les items qui ont un fonctionnement différentiel faible ou très faible.

Comme pour l'ensemble des items qui ont un fonctionnement différentiel non négligeable, le taux de détection des items qui ont un fonctionnement différentiel assez élevé croît en fonction de l'augmentation de la taille des échantillons. L'accroissement des taux suit une trajectoire semblable à celle observée pour la totalité des items qui ont un fonctionnement différentiel non négligeable. La situation n'est pas la même pour les items qui ont un fonctionnement différentiel faible ou très faible. En général, les taux de détection des items qui ont un fonctionnement faible commencent par augmenter, puis diminuent, et ce, pour les critères C_{95} ou $C_{2,5}$ et $C_{97,5}$ et les critères C_{90} ou C_5 et C_{95}. Ils augmentent de façon continue dans le cas des centiles C_{80} ou C_{10} et C_{90}. Les items qui ont un fonctionnement différentiel très faible ont des taux qui diminuent ou se maintiennent sensiblement au même niveau dans le cas des critères C_{95} ou $C_{2,5}$ et $C_{97,5}$ et dans celui des critères C_{90} ou C_5 et C_{95}. Ils tendent à augmenter pour les critères C_{80} ou C_{10} et C_{90}. Autrement dit, il y aurait un effet de plafonnement dans le cas des deux premiers critères. L'augmentation de l'efficacité en fonction de l'accroissement du nombre de sujets se traduit donc par une plus grande capacité de détecter des items qui ont un fonctionnement différentiel plus grand. Par ailleurs, plus le critère de décision permet de considérer comme suspect une quantité importante d'items, plus augmente la possibilité de détecter des items qui ont un fonctionnement différentiel moins élevé. Nous avions noté ces tendances lors de l'examen des fréquences de détection de chaque item.

Nous avons également effectué une analyse de tendance pour l'ensemble des items qui ont un fonctionnement différentiel non négligeable. Pour ce faire, nous avons utilisé le modèle de régression logarithmique. Le tableau 16 reproduit les équations de régression pour chaque critère et chaque méthode à l'étude. Nous y indiquons aussi les coefficients d'ajustement. Les figures 6

à 8 reproduisent les courbes théoriques de régression des taux d'identification correcte en fonction de la grandeur des échantillons. Les taux prédits d'identification correcte sont fournis à l'annexe B (voir les tableaux 9a, 9b et 9c).

Tableau 16. Équations de régression logarithmique des taux d'identification correcte en fonction de la grandeur des échantillons pour des critères basés sur des centiles.

Indice	Critère	Équation	Coefficient R^2
Δ_{MH}	C_1	y = 0,1997ln(x) - 0,5871	0,9797
	C_2	y = 0,1931ln(x) - 0,5904	0,9777
	C_3	y = 0,1557ln(x) - 0,4466	0,9975
DDSM	C_1	y = 0,2109ln(x) - 0,6544	0,9793
	C_2	y = 0,1894ln(x) - 0,5538	0,9966
	C_3	y = 0,1584ln(x) - 0,4486	0,9998
X^2_{MH}	C_1	y = 0,1977ln(x) - 0,5015	0,9606
	C_2	y = 0,2011ln(x) - 0,5908	0,9875
	C_3	y = 0,1723ln(x) - 0,5285	0,9988
$X^2_{amél.}$	C_1	y = 0,2203ln(x) - 0,6968	0,9675
	C_2	y = 0,2052ln(x) - 0,6644	0,9744
	C_3	y = 0,1602ln(x) - 0,4847	0,9951

Légende: C_1 - Critère C_{95} ou $C_{2,5}$ et $C_{97,5}$.
C_2 - Critère C_{90} ou C_5 et C_{95}.
C_3 - Critère C_{80} ou C_{10} et C_{90}.

Les courbes théoriques de régression confirment la plus grande efficacité de la méthode du khi carré de Mantel-Haenszel pour détecter, dans un échantillon particulier, une proportion importante d'items qui ont un fonctionnement différentiel. Par ailleurs, elles situent la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b,

1992 a, 1992 b, 1993 et 1995) au deuxième rang pour les deux critères les plus englobants (C_{90} ou C_5 et C_{95} ; C_{80} ou C_{10} et C_{90}). Elles démontrent aussi la similitude entre cette méthode et la méthode du delta de Mantel-Haenszel. Quels que soient le critère ou la méthode d'analyse du fonctionnement différentiel des items utilisés, les taux prédits diffèrent peu des taux observés.

D'après les courbes théoriques, il faut des échantillons de 400 sujets avec la méthode du khi carré pour espérer que 68 pour cent des items détectés aient effectivement un fonctionnement différentiel lorsque les critères de classification sont les centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$. La méthode de standardisation modifiée exige des échantillons de 575 sujets. Avec les centiles C_{90} ou C_5 et C_{95}, la méthode du khi carré de Mantel-Haenszel exige des échantillons de 575 sujets pour atteindre ce taux d'identification correcte ; la méthode de standardisation modifiée, des échantillons de 670 sujets. Dans le cas des centiles C_{80} ou C_{10} et C_{90}, il faut des échantillons de 1115 sujets avec la méthode du khi carré de Mantel-Haenszel et des échantillons de 1240 sujets avec la méthode de standardisation modifiée.

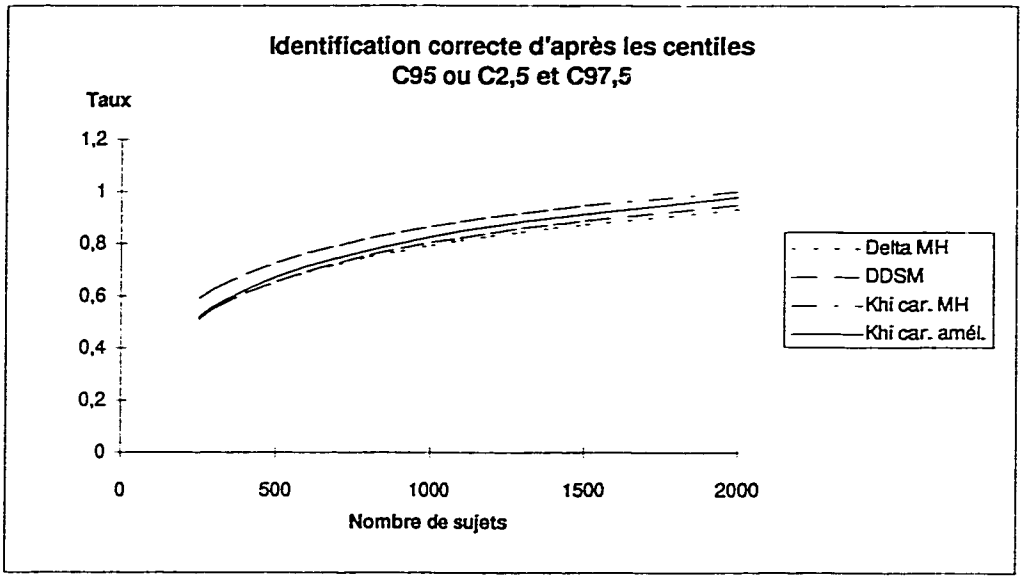


Figure 6. Taux moyens prédits d'identification correcte pour les centiles C95 ou C2,5 et C97,5.

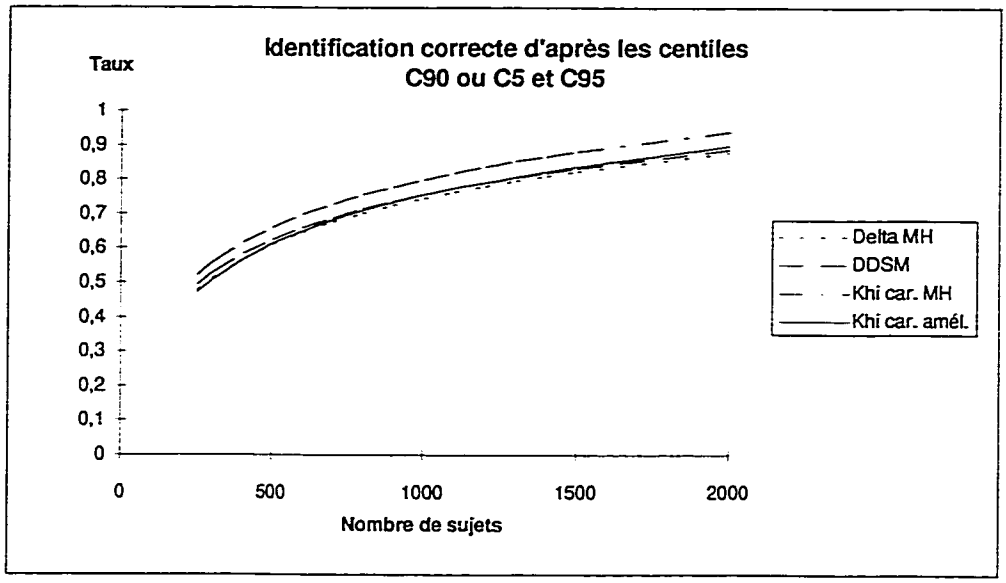


Figure 7. Taux moyens prédits d'identification correcte pour les centiles C90 ou C5 et C95.

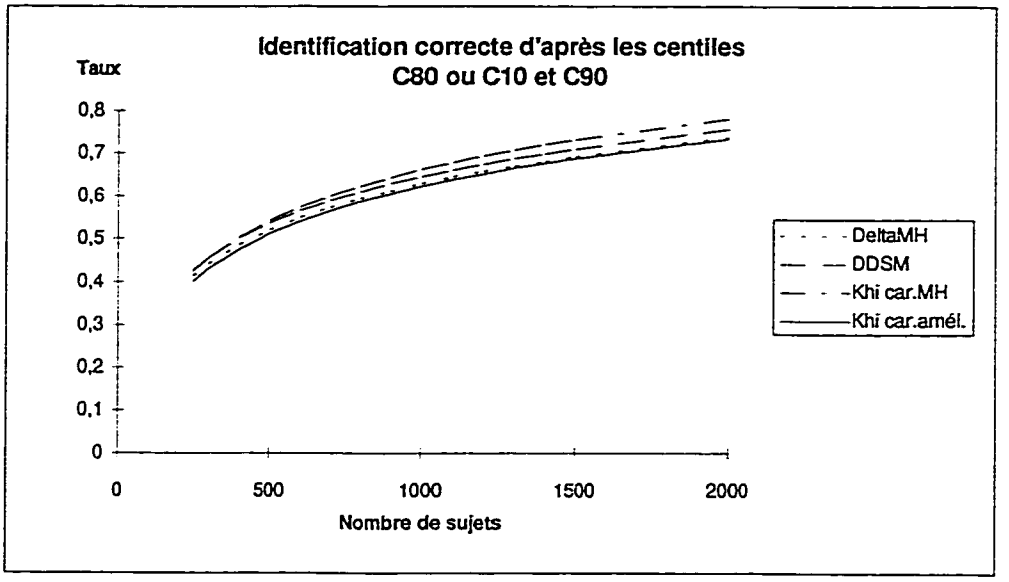


Figure 8. Taux moyens prédits d'identification correcte pour les centiles C80 ou C10 et C90.

Taux relatifs de détection basés sur des critères fixés a priori

La figure 5 (p. 142) reproduit les taux d'identification correcte pour des critères fixés a priori. À titre de rappel, ces critères sont les critères usuels en ce qui concerne la méthode de Mantel-Haenszel et la méthode de régression logistique, à savoir, une valeur de +1,00 ou de -1,00 pour les deltas de Mantel-Haenszel ; la valeur critique du khi carré au seuil de 0,05 à un degré de liberté dans le cas du khi carré de Mantel-Haenszel ; la valeur critique du khi carré au seuil de 0,05 à deux degrés de liberté pour le khi carré d'amélioration. Pour la différence de difficulté standardisée modifiée, nous utilisons la valeur de +0,075 ou -0,075, comme nous l'avons expliqué plus tôt.

À première vue, l'utilisation de critères fixés a priori donne lieu à beaucoup plus de divergence entre les méthodes que le recours à des centiles prédéterminés. De fait, il y a une nette dichotomie entre les méthodes basées sur une mesure de l'ampleur des différences dans le fonctionnement des items (Δ_{MH} et DDSM) et les méthodes basées sur un test statistique (X^2_{MH} et $X^2_{amél.}$). Pour les premières, le taux d'identification correcte croît de façon systématique avec l'augmentation du nombre de sujets dans les échantillons. Pour les secondes, le taux d'identification correcte varie peu. Il y a une très légère augmentation lorsqu'on passe de 250 à 500 sujets, puis de 500 à 1000 sujets, mais une très légère perte lorsqu'on passe de 1000 à 2000 sujets.

Une autre constatation s'impose : la méthode la plus efficace varie en fonction de la taille des groupes. Pour les échantillons de 250 sujets, la méthode la plus efficace est la méthode du khi

carré de Mantel-Haenszel. Elle est suivie de la méthode de régression logistique, puis de la méthode de standardisation modifiée et de la méthode du delta de Mantel-Haenszel. Les deux dernières accusent une différence infime. Pour les échantillons de 500 sujets, la méthode du khi carré de Mantel-Haenszel est la plus efficace. Elle est suivie de la méthode de standardisation modifiée et, sur un pied d'égalité, de la méthode de régression logistique et de la méthode du delta de Mantel-Haenszel. Pour les échantillons de 1000 sujets, la méthode de standardisation modifiée et la méthode du delta de Mantel-Haenszel se partagent la première place. Elles sont suivies de la méthode du khi carré de Mantel-Haenszel et de la méthode de régression logistique. Pour les échantillons de 2000 sujets, la méthode de standardisation modifiée est seule à occuper la première place. Elle est suivie de près par la méthode du delta de Mantel-Haenszel, et de loin par la méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique.

D'après les données recueillies, la méthode de standardisation modifiée deviendrait plus efficace que la méthode du khi carré de Mantel-Haenszel pour des échantillons de plus de 675 sujets et la méthode du delta de Mantel-Haenszel le deviendrait pour des échantillons de plus de 700 sujets. La méthode de régression logistique serait plus efficace que la méthode de standardisation modifiée uniquement pour des échantillons de moins de 475 sujets. Elle serait plus efficace que la méthode du delta de Mantel-Haenszel pour des échantillons de moins de 500 sujets. Elle ne serait jamais supérieure à la méthode du khi carré de Mantel-Haenszel.

Le taux d'identification correcte ne dépasse pas 63 pour cent pour les khis carrés de Mantel-Haenszel et 56 pour cent pour les khis carrés d'amélioration. Avec les deltas de Mantel-Haenszel, le taux d'identification correcte atteint les 68 pour cent avec des échantillons de 860

sujets. Il atteindra les 80 pour cent avec des échantillons de 1350 sujets. Pour les différences de difficulté standardisée modifiée, il faudra des échantillons de 850 sujets pour atteindre un taux de 68 pour cent et des échantillons de 1375 sujets pour le porter à 80 pour cent.

La méthode du delta de Mantel-Haenszel et la méthode de standardisation modifiée peuvent détecter une plus grande proportion d'items qui ont un fonctionnement différentiel non négligeable dans la population que la méthode du khi carré de Mantel-Haenszel ou la méthode de régression logistique. Leur plus grande efficacité doit cependant être interprétée en fonction du nombre total d'items détectés dans les échantillons. À cet égard, le tableau 17 indique le nombre moyen d'items détectés par échantillon pour chaque méthode et chaque grandeur d'échantillon (NTD) et le nombre d'items détectés qui ont un fonctionnement différentiel dans la population (NFD). À la lecture de ce tableau, il appert que l'inefficacité des méthodes basées sur un test statistique tient à leur propension à identifier un nombre d'items qui croît avec l'augmentation du nombre de sujets dans les groupes. Inversement, l'inefficacité des méthodes basées sur une mesure de l'ampleur des différences dans le fonctionnement des items tient à leur propension à identifier un nombre plus grand d'items suspects dans les échantillons plus petits. Autrement dit, l'inefficacité des premières tient à l'utilisation de tests statistiques, lesquelles identifient un nombre croissant d'items avec des différences minimes. L'inefficacité des secondes tient à l'imprécision des indices pour les échantillons de petite taille.

Tableau 17. Nombre moyen d'items détectés dans les échantillons de taille réduite en s'appuyant sur des critères fixés a priori.

Échantillons	Δ_{MH}		DDSM		X^2_{MH}		$X^2_{amél.}$	
	NTD	NFD	NTD	NFD	NTD	NFD	NTD	NFD
250 sujets	17,78	6,63	16,99	6,69	4,82	2,62	6,63	3,07
500 sujets	11,16	5,90	11,10	6,05	8,20	4,92	9,90	5,28
1000 sujets	7,21	5,37	7,35	5,45	13,95	8,79	15,68	8,74
2000 sujets	5,39	4,73	5,30	4,76	21,92	12,90	23,96	12,62

Légende : NTD - Somme du nombre total d'items détectés par échantillon /100 échantillons de même taille.
NFD - Somme du nombre total d'items détectés qui ont un fonctionnement différentiel non négligeable par échantillon/100 échantillons de même taille.

Comme dans le cas des taux d'identification correcte basés sur les centiles, nous avons effectué une analyse de tendance à l'aide du modèle de régression logarithmique. Le tableau 18 fait état des équations de régression et des coefficients d'ajustement obtenus pour chaque méthode et chaque indice à l'étude. La figure 9 reproduit les courbes théoriques de régression des taux d'identification correcte pour chaque méthode et chaque indice à l'étude. Les taux prédits d'identification correcte apparaissent à l'annexe B (voir le tableau 9d).

Tableau 18. Équations de régression logarithmique des taux d'identification correcte en fonction de la grandeur des échantillons pour des critères fixés a priori.

Indice	Critère	Équation	Coefficient R^2
Δ_{MH}	**1,00 ou -1,00**	y = 0,2496ln(x) - 1,0067	0,9923
DDSM	**0,075 ou -0,075**	y = 0,2466ln(x) - 0,9736	0,9975
X^2_{MH}	**0,3841**	y = 0,0238ln(x) + 0,4344	0,3518
$X^2_{amél.}$	**0.5991**	y = 0,0310ln(x) + 0,3166	0,4753

Les courbes théoriques de régression rendent compte de la similitude entre la méthode de standardisation modifiée (DDSM) et la méthode du delta de Mantel-Haenszel. L'écart entre les taux prédits pour chacune des deux méthodes est infime. Il varie entre 1 et 1,9 pour cent selon la grandeur des échantillons. La différence est maximale pour les échantillons de 250 sujets et minimale pour les échantillons de 2000 sujets. À partir des valeurs prédites, la méthode de standardisation modifiée présente les taux les plus élevés. Les taux prédits pour les deux méthodes diffèrent peu des taux observés. Les taux prédits pour la méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique s'écartent davantage des taux observés. De fait, un modèle polynomial donne un meilleur ajustement, mais des valeurs prédites s'avèrent supérieures à l'unité et le modèle s'ajuste mal aux données des deltas de Mantel-Haenszel et des différences de difficulté standardisée modifiée.

À partir des valeurs prédites, la méthode de standardisation modifiée nécessiterait des échantillons de 820 sujets pour que 68 pour cent des items détectés aient effectivement un fonctionnement différentiel non négligeable dans la population. La méthode du delta de Mantel-Haenszel exigerait des échantillons de 850 sujets. Pour atteindre les 80 pour cent, la première demanderait des échantillons de 1320 sujets et la deuxième, des échantillons de 1390 sujets.

L'examen des taux de détection dans le tableau 15d (p. 138) montre que les taux de détection des items qui ont un fonctionnement différentiel assez élevé suit la même trajectoire que celle observée pour la totalité des items avec un fonctionnement différentiel non négligeable. Par contre, les items qui ont un fonctionnement différentiel faible ou très faible ont des taux de détection qui se comportent différemment. Si nous ne considérions comme ayant un fonctionnement différentiel non négligeable que les items qui accusent des différences assez élevées, nous observerions donc sensiblement les mêmes tendances que celle observées pour les 15 items retenus. Le taux d'identification correcte serait toutefois plus petit, et le taux d'identification incorrecte serait accru de la proportion d'items qui affichent un fonctionnement différentiel faible ou très faible.

L'examen des taux de détection des indices orientés (Δ_{MH} et DDSM) montre également que la détection des items qui ont un fonctionnement différentiel plus grand augmente avec l'accroissement du nombre de sujets dans les groupes. Par contre, la détection des items qui ont un fonctionnement différentiel faible ou très faible n'augmente pas ou tout au moins, pas autant. Pour les indices non orientés (X_{MH} et $X_{amél.}$), c'est l'inverse qui se produit. C'est la proportion d'items avec un fonctionnement différentiel faible ou très faible qui augmente, une autre manifestation de l'effet de la grandeur des échantillons sur la valeur calculée des indices.

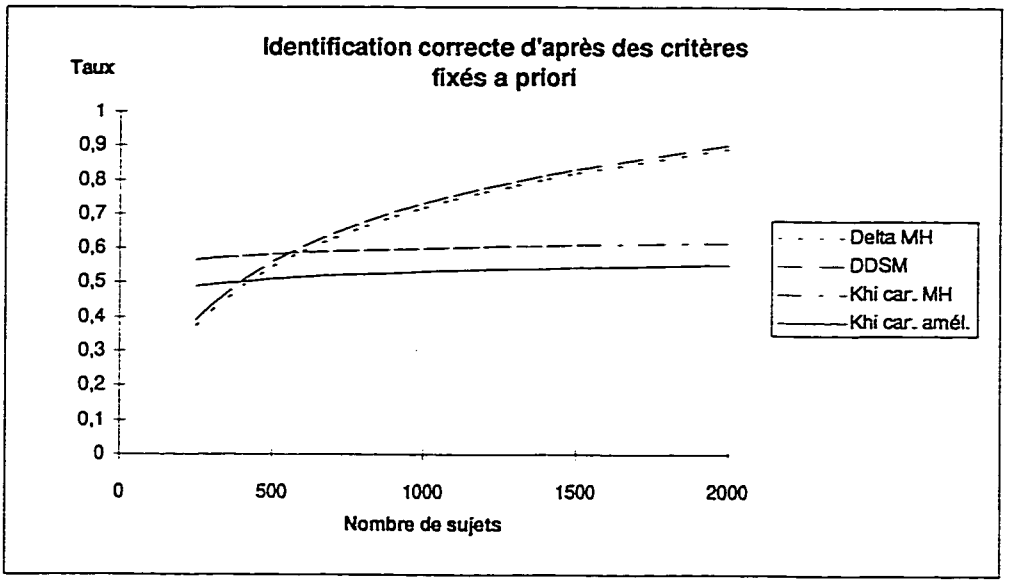


Figure 9. Taux moyens prédits d'identification correcte pour des critères fixés a priori.

Les taux objectifs de détection

Les taux objectifs de détection consistent en des taux calculés en fonction des items qui ont ou qui n'ont pas de fonctionnement différentiel (FD) dans la population, sans égard au nombre total d'items détectés dans les échantillons. Ces taux présupposent une répartition des données dans un tableau de contingence à deux dimensions et à quatre cases. Ils renvoient alors aux notions de vrais positifs (VP), de faux positifs (FP), de faux négatifs (FN) et de vrais négatifs (VN), selon que les items détectés dans les échantillons ont effectivement un fonctionnement différentiel non négligeable (items avec FD) ou un fonctionnement différentiel négligeable (items sans FD) dans la population.

		Items dans la population		
		avec FD	sans FD	
Items dans les échantillons	détectés	Vrais positifs (VP)	Faux positifs (FP)	Nombre total d'items détectés
	non détectés	Faux négatifs (FN)	Vrais négatifs (VN)	Nombre total d'items non détectés
		15 items	45 items	

Dans ce contexte, le taux de vrais positifs correspond à la proportion moyenne d'items détectés qui ont effectivement un fonctionnement différentiel par rapport aux 15 items identifiés comme

ayant un fonctionnement différentiel non négligeable dans la population, pour les 100 échantillons de même taille.

$$Taux_{VP} = \sum [\frac{Nombre_{VP}}{15}] \ / \ 100 \qquad (28)$$

Le taux de faux positifs correspond à la proportion moyenne d'items détectés, identifiés à tort comme ayant un fonctionnement différentiel, par rapport aux 45 items qui ont un fonctionnement différentiel négligeable dans la population, pour les 100 échantillons de même taille.

$$Taux_{FP} = \sum [\frac{Nombre_{FP}}{45}] \ / \ 100 \qquad (29)$$

À partir des données à disposition, il est possible de déterminer le taux de succès. Pour ce faire, il suffit de déterminer le taux de vrais négatifs, puis de faire la somme pondérée du taux de vrais positifs et du taux de vrais négatifs. Le taux de vrais négatifs est donné par la différence entre le taux de faux positifs et l'unité.

$$Taux_{succès} = (Taux_{VP})(0,25) + (Taux_{VN})(0,75) \qquad (30)$$

où

$$Taux_{VN} = 1 - Taux_{FP} \qquad (31)$$

Ce taux devrait refléter le nombre moyen de décisions correctes estimé précédemment (voir les tableaux 14a, 14b, 14c et 14d aux pages 128 à 131). Nous n'aborderons pas la question ici.

Les tableaux 19a, 19b, 19c et 19d reproduisent les taux de vrais et de faux positifs obtenus pour chaque critère et chaque grandeur d'échantillon utilisés. Les taux de vrais positifs correspondent aux taux d'items détectés qui ont un fonctionnement différentiel (FD) non négligeable dans la population. Les taux de faux positifs correspondent aux taux d'items détectés qui ont un fonctionnement différentiel (FD) négligeable dans cette même population. Pour les items qui ont un fonctionnement différentiel non négligeable, nous indiquons aussi les taux de vrais positifs si nous n'avions considéré comme ayant un fonctionnement différentiel non négligeable que les cinq items qui ont un fonctionnement différentiel assez élevé, faible ou très faible, ceci pour connaître la grandeur exacte des taux de vrais positifs dans de telles conditions. Le tableau 7 (p. 81) précise les caractéristiques des trois groupes d'items. En ce qui concerne les taux fournis dans les tableaux 19a, 19b, 19c et 19d, le nombre d'items détectés dans les échantillons a été divisé par cinq au lieu de 15. Pour connaître la répartition des 15 items en fonction de l'ampleur des différences dans le fonctionnement des items, il faudra donc diviser les taux indiqués par trois.

Les figures 10, 11, 12 et 13 reproduisent les taux de vrais positifs et de faux positifs en fonction de la taille des échantillons. Les figures 10, 11 et 12 font état des taux établis à partir de centiles prédéterminés comme critère de décision. La figure 13 reproduit les taux de vrais positifs et de faux positifs en fonction de critères fixés a priori. Règle générale, les taux de vrais positifs sont inférieurs aux taux d'identification correcte (voir les tableaux 15a, 15b, 15c et 15d aux pages 135 à 138). Cette diminution est essentiellement due à la prise en considération de la totalité des items détectés pour le calcul des taux relatifs. Nous examinerons d'abord les taux basés sur des centiles prédéterminés, puis ceux qui sont basés sur des critères fixés a priori.

Tableau 19a. Taux moyens de vrais positifs et taux moyens de faux positifs en fonction de la grandeur des échantillons en appliquant les centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$ comme critère de décision (N = 100 échantillons).

Indice	FD	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}	**Négligeable**	0,04467	0,03022	0,01578	0,00844
	Non négligeable	**0,13267**	**0,18000**	**0,22000**	**0,24467**
	Très faible	0,08000	0,06400	0,06800	0,05400
	Faible	0,10600	0,15800	0,15800	0,14000
	Assez élevé	0,21200	0,31800	0,43400	0,54000
DDSM	**Négligeable**	0,04733	0,02822	0,01644	0,00644
	Non négligeable	**0,13333**	**0,18667**	**0,21933**	**0,25200**
	Très faible	0,11000	0,07800	0,10200	0,08800
	Faible	0,09400	0,16800	0,15000	0,14000
	Assez élevé	0,19600	0,31400	0,40600	0,52800
X^2_{MH}	**Négligeable**	0,02933	0,01622	0,06889	0,00200
	Non négligeable	**0,11200**	**0,15133**	**0,18067**	**0,19400**
	Très faible	0,06800	0,04200	0,04600	0,03000
	Faible	0,08800	0,14600	0,12200	0,09200
	Assez élevé	0,18000	0,26600	0,37400	0,46000
$X^2_{amél.}$	**Négligeable**	0,03378	0,02067	0,00889	0,00378
	Non négligeable	**0,09867**	**0,13800**	**0,17333**	**0,18867**
	Très faible	0,06800	0,05400	0,04800	0,02800
	Faible	0,07800	0,11600	0,11200	0,10800
	Assez élevé	0,15000	0,24000	0,36000	0,43000

NB: Les taux en caractères gras correspondent aux taux de vrais positifs pour la totalité des items identifiés comme ayant un fonctionnement différentie(FD). Sous ces taux, on indique le taux de vrais positifs pour des groupes restreints d'items. Au-dessus, apparaissent les taux de faux positifs pour la totalité des items qui n'ont pas de fonctionnement différentiel (FD).

Tableau 19b. Taux moyens de vrais positifs et taux moyens de faux positifs en fonction de la grandeur des échantillons en appliquant les centiles C_{90} ou C_5 et C_{95} comme critère de décision (N = 100 échantillons).

Indice	FD	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}	**Négligeable**	0,07356	0,05044	0,03133	0,02000
	Non négligeable	**0,18400**	**0,25467**	**0,31467**	**0,34867**
	Très faible	0,12200	0,11800	0,13800	0,11600
	Faible	0,14000	0,23800	0,25400	0,24600
	Assez élevé	0,29000	0,40800	0,55200	0,68400
DDSM	**Négligeable**	0,06978	0,04956	0,03289	0,01644
	Non négligeable	**0,19533**	**0,25933**	**0,30933**	**0,36000**
	Très faible	0,16800	0,13400	0,14200	0,17800
	Faible	0,15000	0,23600	0,24400	0,24400
	Assez élevé	0,26800	0,40800	0,54200	0,65800
X^2_{MH}	**Négligeable**	0,06667	0,04333	0,02444	0,01067
	Non négligeable	**0,20200**	**0,27067**	**0,32733**	**0,36800**
	Très faible	0,13400	0,11600	0,13400	0,12400
	Faible	0,18000	0,26800	0,27800	0,27400
	Assez élevé	0,29200	0,42800	0,57000	0,70600
$X^2_{amél.}$	**Négligeable**	0,07356	0,05133	0,02800	0,01800
	Non négligeable	**0,18000**	**0,24600**	**0,31600**	**0,34600**
	Très faible	0,11800	0,11200	0,13000	0,11000
	Faible	0,14800	0,22000	0,26000	0,24800
	Assez élevé	0,27400	0,40600	0,55800	0,68000

NB: Les taux en caractères gras correspondent aux taux de vrais positifs pour la totalité des items identifiés comme ayant un fonctionnement différentiel (FD). Sous ces taux, on indique le taux de vrais positifs pour des groupes restreints d'items. Au-dessus, apparaissent les taux de faux positifs pour la totalité des items qui n'ont pas de fonctionnement différentiel (FD).

Tableau 19c. Taux moyens de vrais positifs et taux moyens de faux positifs en fonction de la grandeur des échantillons en appliquant les centiles C_{80} ou C_{10} et C_{90} comme critère de décision (N = 100 échantillons).

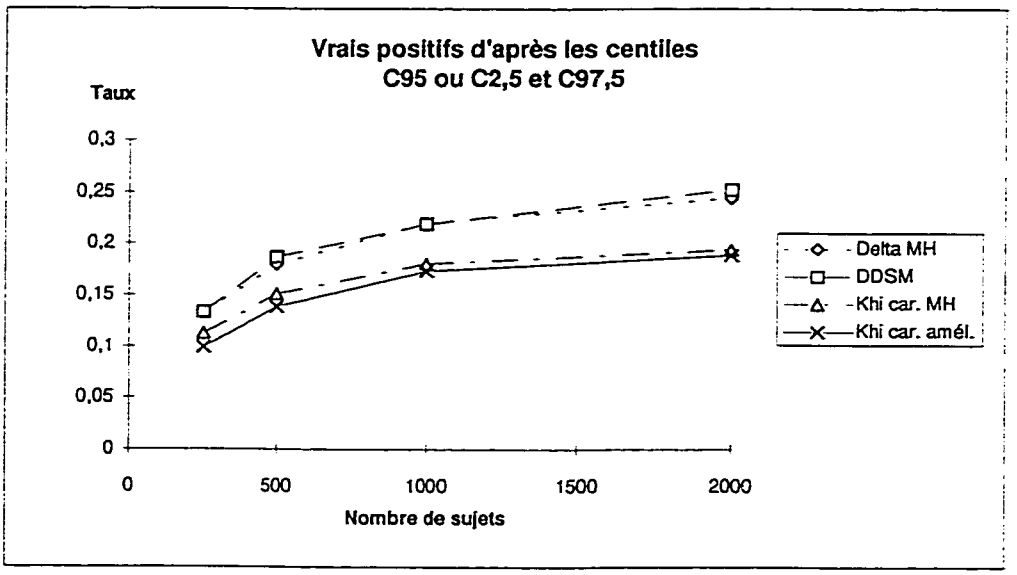
Indice	FD	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}	**Négligeable**	0,15911	0,12667	0,09933	0,07267
	Non négligeable	**0,32733**	**0,42733**	**0,51333**	**0,59600**
	Très faible	0,25600	0,26400	0,30800	0,36600
	Faible	0,29000	0,41200	0,48000	0,56200
	Assez élevé	0,43600	0,60600	0,75200	0,86000
DDSM	**Négligeable**	0,15489	0,12644	0,95111	0,06689
	Non négligeable	**0,34533**	**0,43467**	**0,52533**	**0,61600**
	Très faible	0,28200	0,28800	0,34000	0,43400
	Faible	0,31000	0,42000	0,49600	0,57000
	Assez élevé	0,44400	0,59600	0,74000	0,84400
X^2_{MH}	**Négligeable**	0,15467	0,12244	0,08822	0,05956
	Non négligeable	**0,33733**	**0,43467**	**0,53600**	**0,62133**
	Très faible	0,26000	0,28000	0,31800	0,39000
	Faible	0,30200	0,43200	0,51200	0,58400
	Assez élevé	0,45000	0,59200	0,77800	0,89000
$X^2_{amél.}$	**Négligeable**	0,16156	0,13000	0,09733	0,07378
	Non négligeable	**0,31533**	**0,41000**	**0,50867**	**0,57867**
	Très faible	0,24000	0,24600	0,30000	0,34200
	Faible	0,28000	0,40800	0,47400	0,52400
	Assez élevé	0,42600	0,57600	0,75200	0,87000

NB: Les taux en caractères gras correspondent aux taux de vrais positifs pour la totalité des items identifiés comme ayant un fonctionnement différentiel (FD). Sous ces taux, on indique le taux de vrais positifs pour des groupes restreints d'items. Au-dessus, apparaissent les taux de faux positifs pour la totalité des items qui n'ont pas de fonctionnement différentiel (FD).

Tableau 19d. Taux moyens de vrais positifs et taux moyens de faux positifs en fonction de la grandeur des échantillons en appliquant des critères fixés a priori comme critère de décision (N = 100 échantillons).

Indice	FD	250 sujets	500 sujets	1000 sujets	2000 sujets
Δ_{MH}	**Négligeable**	0,24778	0,11689	0,04089	0,01467
	Non négligeable	**0,44200**	**0,39333**	**0,35800**	**0,31533**
	Très faible	0,38200	0,24200	0,17400	0,10800
	Faible	0,39400	0,37800	0,29400	0,20400
	Assez élevé	0,55000	0,56000	0,60600	0,63400
DDSM	**Négligeable**	0,22889	0,11222	0,04222	0,01200
	Non négligeable	**0,44600**	**0,40333**	**0,36333**	**0,31733**
	Très faible	0,38800	0,26600	0,19200	0,13600
	Faible	0,40200	0,39000	0,30800	0,21800
	Assez élevé	0,54800	0,55400	0,59000	0,59800
X^2_{MH}	**Négligeable**	0,04889	0,07289	0,11467	0,20044
	Non négligeable	**0,17467**	**0,32800**	**0,58600**	**0,86000**
	Très faible	0,10800	0,16400	0,40200	0,72000
	Faible	0,15600	0,32200	0,55600	0,88600
	Assez élevé	0,26000	0,49800	0,80000	0,97400
$X^2_{amél.}$	**Négligeable**	0,07911	0,10267	0,15422	0,25200
	Non négligeable	**0,20467**	**0,35200**	**0,58267**	**0,84133**
	Très faible	0,15200	0,19000	0,39200	0,70200
	Faible	0,16800	0,34000	0,56400	0,85800
	Assez élevé	0,29400	0,52600	0,79200	0,96400

NB: Les taux en caractères gras correspondent aux taux de vrais positifs pour la totalité des items identifiés comme ayant un fonctionnement différentie (FD). Sous ces taux, on indique le taux de vrais positifs pour des groupes restreints d'items. Au-dessus, apparaissent les taux de faux positifs pour la totalité des items qui n'ont pas de fonctionnement différentiel (FD).



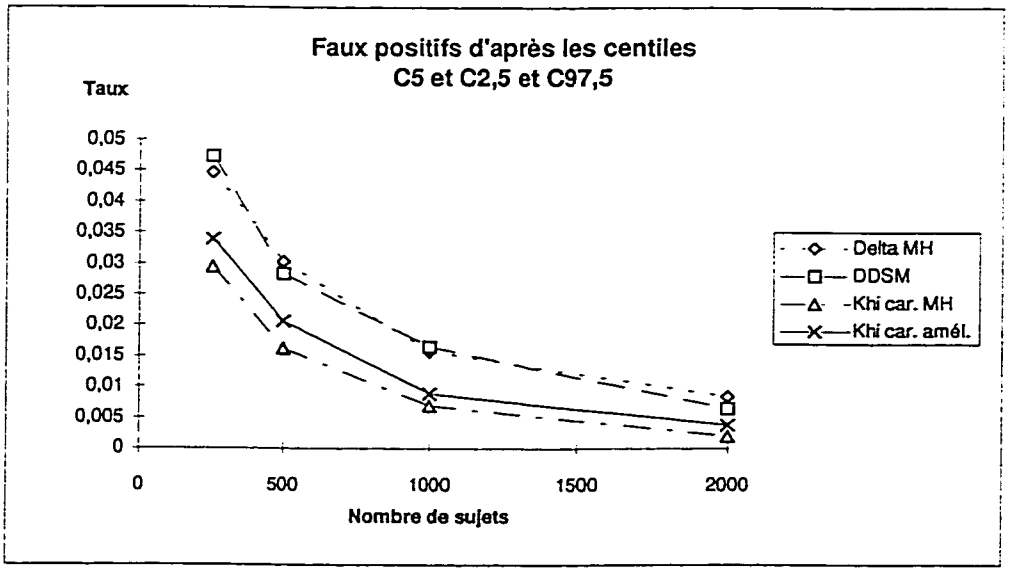
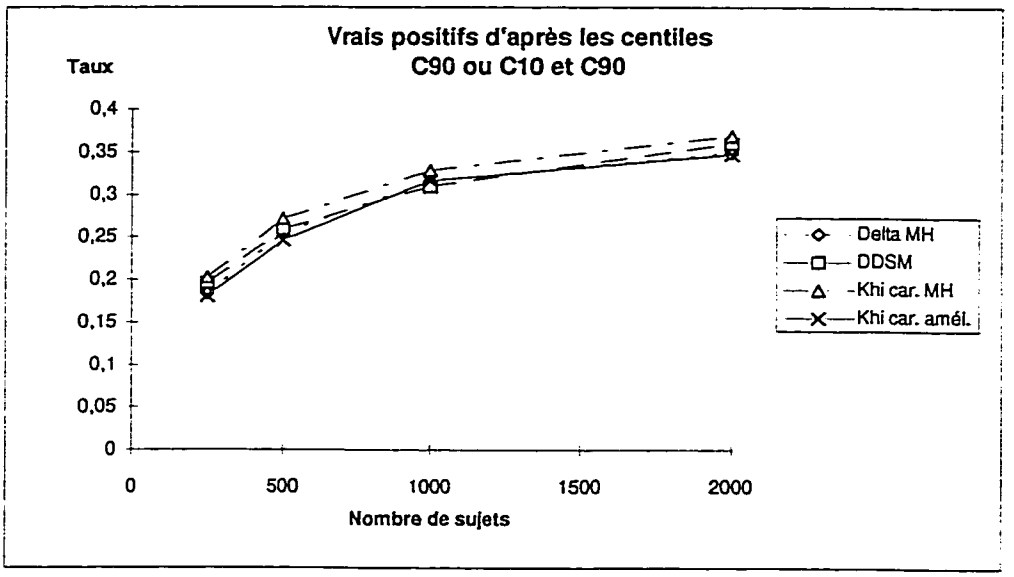


Figure 10. Taux moyens observés de vrais ou de faux positifs pour les critères C95 ou C2,5 et C97,5.



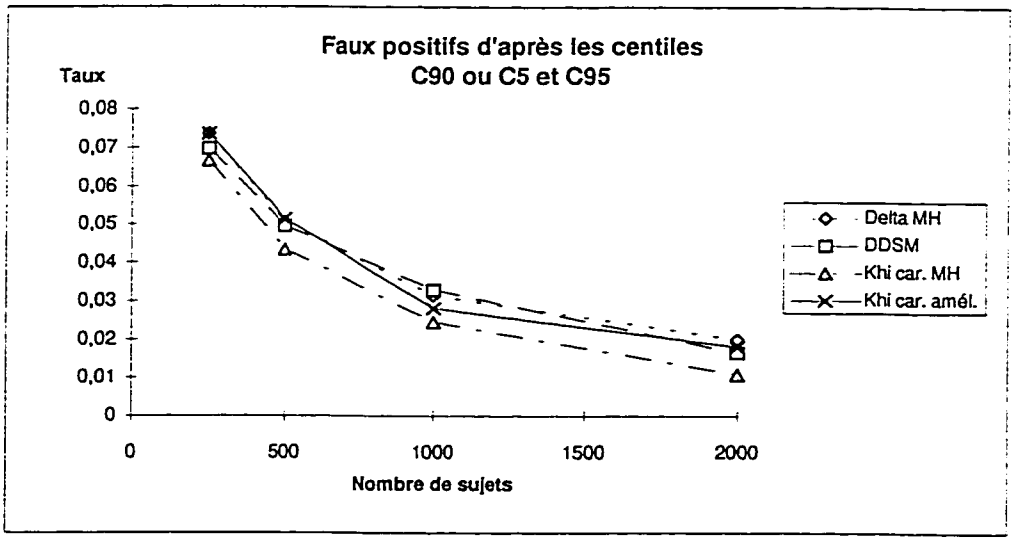
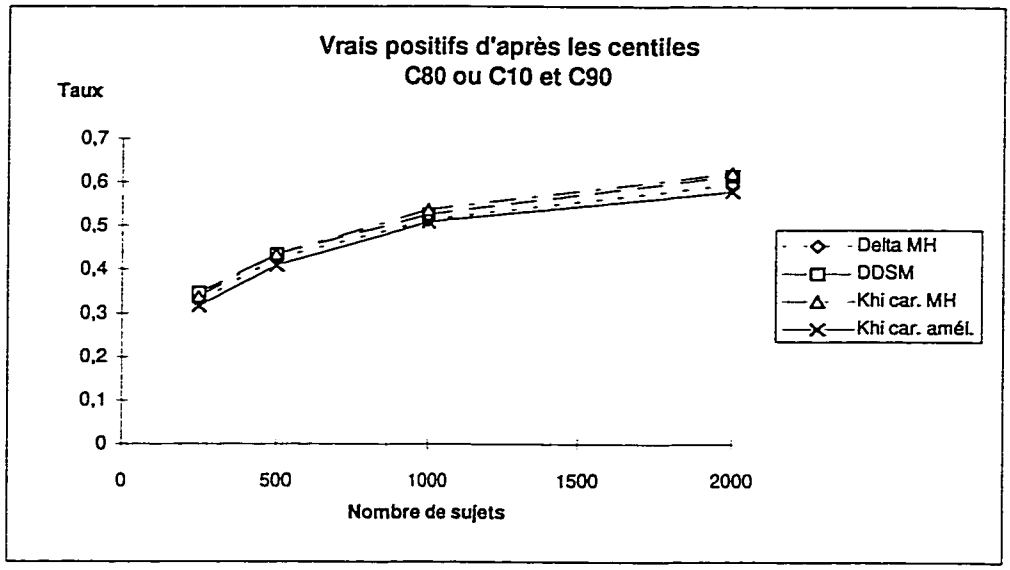


Figure 11. Taux moyens observés de vrais ou de faux positifs pour les critères C90 ou C5 et C95.



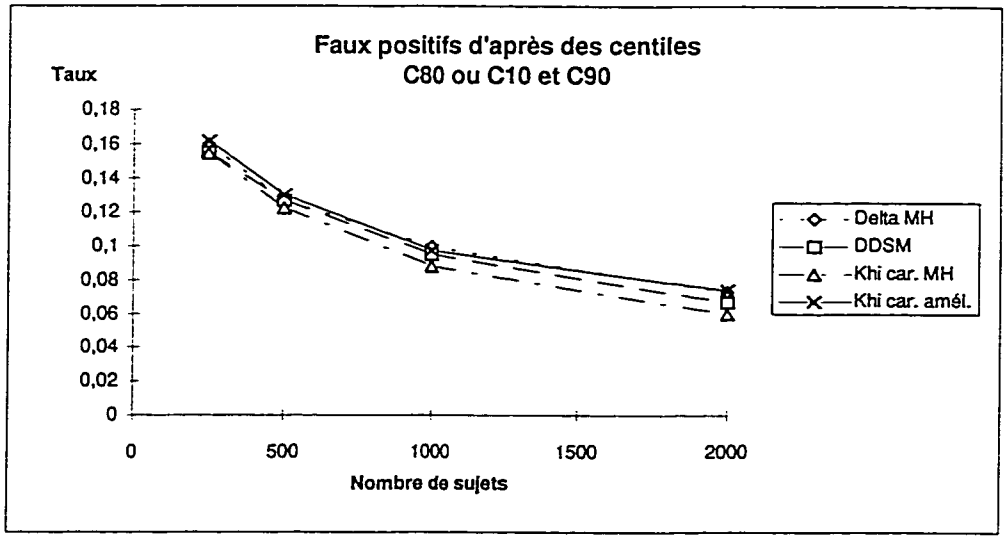
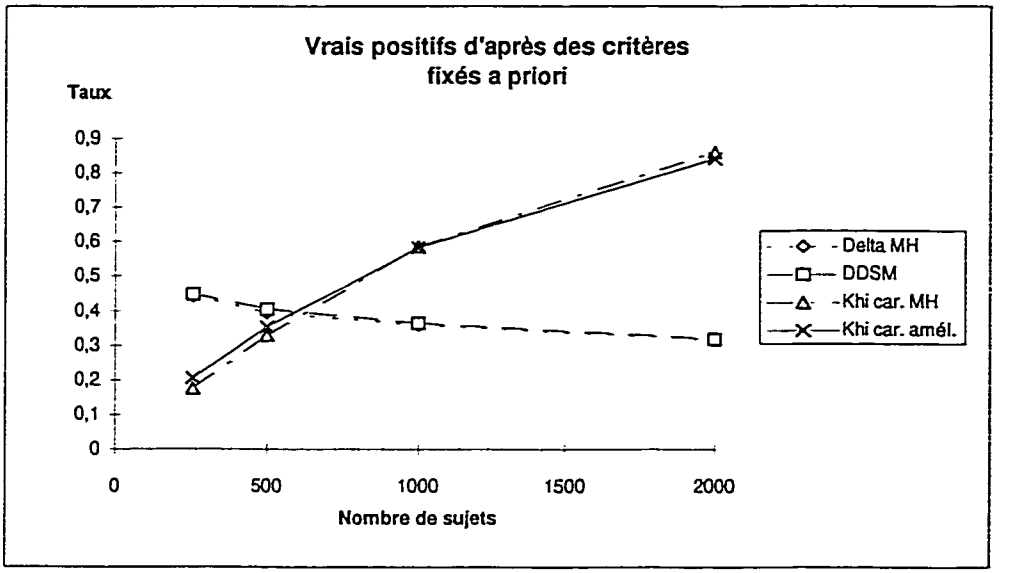


Figure 12. Taux moyens observés de vrais ou de faux positifs pour les critères C80 ou C10 et C90.



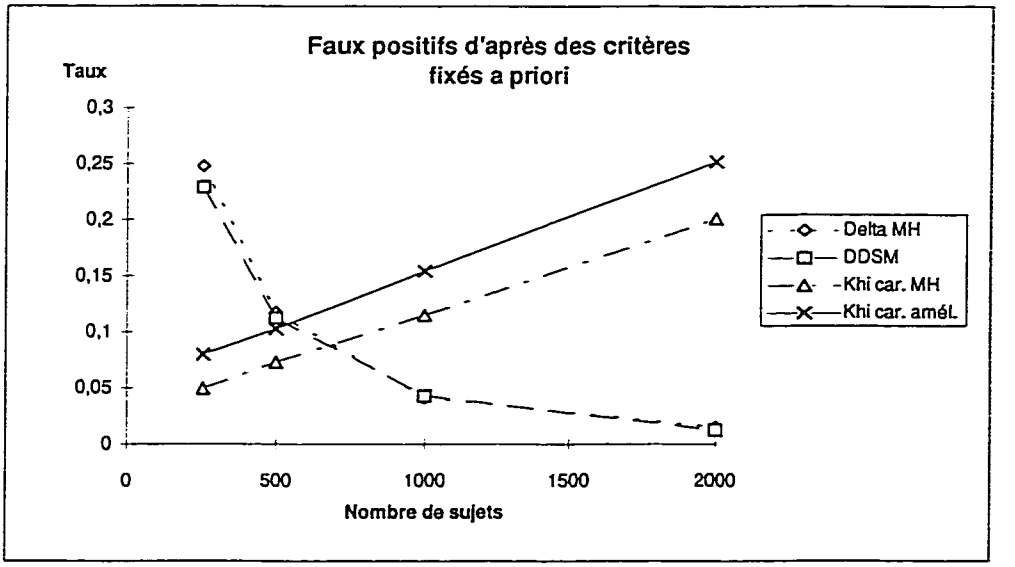


Figure 13. Taux moyens observés de vrais ou de faux positifs pour des critères fixés a priori.

Taux objectifs basés sur des centiles prédéterminés

L'examen des tableaux 19a, 19b et 19c et l'examen des figures 10, 11, et 12 montrent que les taux de vrais positifs augmentent en fonction de la taille des échantillons et que les taux de faux positifs diminuent en conséquence, un phénomène prévisible et attendu. Ils montrent aussi que les taux de vrais positifs sont plus grands pour une même grandeur d'échantillon lorsque le critère de décision est moins exigeant (les centiles C_{80} ou C_{10} et C_{90}). L'augmentation du taux de vrais positifs s'accompagne d'une augmentation du taux de faux positifs. Plus d'items sont détectés. Parmi ces derniers, il y a plus de vrais positifs et, également, plus de faux négatifs. En principe, les taux de vrais positifs devraient refléter les tendances observées avec les taux relatifs, puisque le critère de décision dépend uniquement de la grandeur des indices de fonctionnement différentiel. Dans le cas présent, les données confirment la supériorité de la méthode du khi carré de Mantel-Haenszel pour les centiles C_{90} ou C_5 et C_{95} et les centiles C_{80} ou C_{10} et C_{90}. Il n'en va pas de même pour les centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$. Ce sont la méthode de standardisation modifiée et la méthode du delta de Mantel-Haenszel qui s'avèrent les plus efficaces. Le renversement de situation par rapport aux taux relatifs s'explique par le fait que quatre items sont identifiés comme ayant un fonctionnement différentiel dans le cas des indices orientés (Δ_{MH} et DDSM) et trois dans le cas des indices non orientés (X^2_{MH} et $X^2_{amél.}$). Les taux relatifs ne prennent pas ce facteur en considération.

En général, les taux de vrais positifs sont faibles, en tout cas plus faibles que nous l'aurions cru. Pour la méthode du delta de Mantel-Haenszel et les centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$, ils varient de 13,3 à 24,5 pour cent selon la grandeur des échantillons. Pour la méthode de standardisation

modifiée, ils vont de 13,3 à 25,2 pour cent. Pour la méthode du khi carré de Mantel-Haenszel, ils passent de 11,2 à 19,4 pour cent selon la grandeur des échantillons ; et pour la méthode de régression logistique, de 9.9 à 18,9 pour cent.

Pour les centiles C_{90} ou C_5 et C_{95}, la méthode du delta de Mantel-Haenszel offre des taux qui varient de 18,4 à 34,9 pour cent ; et la méthode de standardisation modifiée, des taux qui vont de 19,5 à 36,0 pour cent. Les taux correspondants vont de 20,2 à 36,8 pour la méthode du khi carré de Mantel-Haenszel et de 18,0 à 34,6 pour cent pour la méthode de régression logistique.

En ce qui concerne les centiles C_{80} ou C_{10} et C_{90}, les taux de vrais positifs varient de 32,7 à 59,6 pour cent pour la méthode du delta de Mantel-Haenszel ; et de 34,5 à 61,6 pour cent pour la méthode de standardisation modifiée. Dans le cas de la méthode du khi carré de Mantel-Haenszel, ils passent de 33,7 à 62,1 pour cent ; et dans le cas de la méthode de régression logistique, de 31,5 à 57,9 pour cent.

La faiblesse des taux de vrais positifs doit être interprétée avec circonspection. Nous avons identifié 15 items comme ayant un fonctionnement différentiel non négligeable dans la population. Ces derniers représentent 25 pour cent de tous les items contenus dans le test. Le choix des centiles utilisés comme critère de décision se situe dans une perspective pragmatique, celle du praticien qui ne sait pas combien d'items ont un fonctionnement différentiel et qui décide de retenir comme suspects cinq, 10 ou 20 pour cent des items analysés. Aucun de ces centiles ne permet de détecter la totalité des items qui ont un fonctionnement différentiel non négligeable. Les centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$ permettent d'identifier quatre items dans le cas des

indices orientés (Δ_{MH} et DDSM) et trois dans celui des indices non orientés (X^2_{MH} et $X^2_{amél.}$). Le taux maximal possible de vrais positifs est alors de 26,7 pour cent pour les indices orientés et de 20,0 pour cent pour les indices non orientés. Avec les centiles C_{90} ou C_5 et C_{95}, six items sont détectés. Par conséquent, le taux maximal possible de vrais positifs est de 40 pour cent. Avec les centiles C_{80} ou C_{10} et C_{90}, 12 items sont détectés et le taux maximal possible de vrais positifs s'élève à 80 pour cent. Quels que soient les centiles utilisés comme critère de décision, le taux maximal possible de vrais positifs n'est jamais atteint. Dans le cas des centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$, le taux maximal est presqu'atteint avec la méthode du khi carré de Mantel-Haenszel et des échantillons de 2000 sujets. Il se situe entre 1,1 et 2,2 pour cent du maximum avec les autres méthodes (la méthode du delta de Mantel-Haenszel, la méthode de standardisation modifiée et la méthode de régression logistique). Pour les centiles C_{90} ou C_5 et C_{95}, il se situe entre 3,2 et 5,4 pour cent du maximum, selon les méthodes et les indices considérés. Pour les centiles C_{80} ou C_{10} et C_{90}, il se situe entre 17,9 et 22,1 pour cent du maximum possible.

Comme dans le cas des taux relatifs d'identification correcte, nous avons effectué des analyses de tendance à l'aide du modèle de régression logarithmique. Le tableau 20 reproduit les équations de régression pour chacune des méthodes à l'étude et les coefficients d'ajustement. Les figures 14, 15 et 16 reproduisent les courbes théoriques obtenues à partir de ces équations. Les taux prédits de vrais positifs figurent à l'annexe B (voir les tableaux 10a, 10b et 10c).

Tableau 20. Équations de régression logarithmique des taux de vrais positifs en fonction de la grandeur des échantillons pour des critères basés sur des centiles.

Indice	Critère	Équation	Coefficient R^2
Δ_{MH}	C_1	y = 0,0542ln(x) - 0,1616	0,9817
	C_2	y = 0,0799ln(x) - 0,2489	0,9778
	C_3	y = 0,1287ln(x) - 0,3783	0,9980
DDSM	C_1	y = 0,0561ln(x) - 0,1701	0,9833
	C_2	y = 0,0785ln(x) - 0,2339	0,9963
	C_3	y = 0,1302ln(x) - 0,3741	1,0000
X^2_{MH}	C_1	y = 0,0397ln(x) - 0,1011	0,9569
	C_2	y = 0,0800ln(x) - 0,2330	0,9874
	C_3	y = 0,1375ln(x) - 0,4201	0,9988
$X^2_{amél.}$	C_1	y = 0,0441ln(x) - 0,1394	0,9675
	C_2	y = 0,0819ln(x) - 0,2657	0,9746
	C_3	y = 0,1282ln(x) - 0,3880	0,9948

Légende: C_1 - Critère C_{95} ou $C_{2,5}$ et $C_{97,5}$.
C_2 - Critère C_{90} ou C_5 et C_{95}.
C_3 - Critère C_{80} ou C_{10} et C_{90}.

Les courbes théoriques de régression reflètent les tendances constatées à partir des taux observés. L'écart entre les taux observés et les taux prédits est plus petit que celui qui prévaut dans le cas des taux relatifs. Il ne dépasse pas le seuil d'un pour cent, mis à part le cas des khis carrés d'amélioration avec les échantillons de 2000 sujets et les centiles C_{90} ou C_5 et C_{95} comme critère de décision. Elles confirment aussi la similitude entre la méthode de standardisation modifiée et la méthode du delta de Mantel-Haenszel et le peu de différence entre les diverses méthodes pour ce qui est du taux de vrais positifs pour des échantillons de même grandeur. Vue sous cet angle, la méthode de standardisation modifiée n'est pas supérieure aux autres, mais elle n'est pas

inférieure. De fait, les analyses de tendance la placent au premier rang, avant la méthode du delta de Mantel-Haenszel pour les centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$. Elles la mettent au deuxième rang, après la méthode du khi carré de Mantel-Haenszel, pour les centiles C_{90} ou C_5 et C_{95}. Enfin, pour ce qui est des centiles C_{80} ou C_{10} et C_{90}, elles la situent au premier rang pour les échantillons de 250 sujets et au deuxième pour les échantillons de 1000 et de 2000 sujets.

Les analyses de tendance effectuées indiquent qu'il n'y a que 25,6 pour cent de vrais positifs détectés lorsqu'on se sert des centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$ comme critère de décision. Ce taux est atteint par la méthode de standardisation modifiée avec des échantillons de 2000 sujets. La méthode du delta de Mantel-Haenszel suit de près avec un taux de 25,0 pour cent. Avec la même grandeur d'échantillon, le taux de vrais positifs s'élève à 20,1 pour cent pour la méthode du khi carré de Mantel-Haenszel et à 19,6 pour cent pour la méthode de régression logistique. Avec des échantillons de 2000 sujets et les centiles C_{90} ou C_5 et C_{95} comme critère de décision, on ne peut espérer détecter plus de 37,5 pour cent de vrais positifs, et ce, avec la méthode du khi carré de Mantel-Haenszel, la plus efficace dans les conditions précitées. Le taux de vrais positifs varie de 35,7 à 36,3 pour cent avec les autres méthodes (35,8 pour cent dans le cas des deltas de Mantel-Haenszel, 36,3 pour cent dans le cas des différences de difficulté standardisée modifiée et 35,7 pour cent dans le cas de la méthode de régression logistique). Pour les centiles C_{80} ou C_{10} et C_{90}, le taux maximal de vrais positifs s'élève à 62,5 pour cent pour des échantillons de 2000 sujets et la méthode du khi carré de Mantel-Haenszel. Avec la même grandeur d'échantillon, la méthode de standardisation modifiée affiche un taux de 61,6 pour cent ; la méthode du delta de Mantel-Haenszel, un taux de 60,0 pour cent ; et la méthode de régression logistique, un taux de 58,6 pour cent.

La faiblesse des taux de vrais positifs dépend non seulement des centiles utilisés comme critère de décision mais aussi de la faiblesse des différences dans le fonctionnement des items qui présentent un fonctionnement différentiel non négligeable. La différence de difficulté standardisée (DDS) observée dans la population présente une moyenne de 0,06507 en valeur absolue, pour les 15 items retenus. Les items qui présentent un fonctionnement différentiel assez élevé affichent une moyenne de 0,08340 en valeur absolue. Si nous considérions que seuls ces items ont un fonctionnement différentiel non négligeable, les taux de vrais positifs seraient nettement plus grands. À l'examen des tableaux 19a, 19b et 19c (p. 164 à 166), ils se situent entre 21,2 et 54,0 pour cent, selon la grandeur des échantillons, pour la méthode du delta de Mantel-Haenszel et les centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$; entre 29,0 et 68,4 pour cent pour les centiles C_{90} ou C_5 et C_{95} ; entre 43,6 et 86,0 pour cent pour les centiles C_{80} ou C_{10} et C_{90}. Nous observons une augmentation comparable pour les autres méthodes à l'étude. Bref, tout porte à croire que les taux de vrais positifs, et partant l'efficacité des méthodes, seraient supérieurs. Par contre, les taux de faux positifs augmenteraient en conséquence. Par ailleurs, le taux maximal possible de vrais positifs serait de 100 pour cent avec les centiles C_{90} ou C_5 et C_{95} et les centiles C_{80} ou C_{10} et C_{90} comme critère de décision.

Présentement, la méthode de standardisation modifiée tend à se situer au deuxième rang. Lorsqu'on ne considère que les cinq items avec un fonctionnement différentiel assez élevé, elle suit une trajectoire qui la situe au deuxième rang pour les centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$ et au quatrième pour les centiles C_{90} ou C_5 et C_{95}. Pour les centiles C_{80} ou C_{10} et C_{90}, elle passe du deuxième rang pour les échantillons de 250 et de 500 sujets au quatrième pour les échantillons de 1000 ou de 2000 sujets. Vue sous cet angle, sa supériorité présente serait due à sa plus grande capacité de détecter des items qui ont un fonctionnement différentiel plus faible.

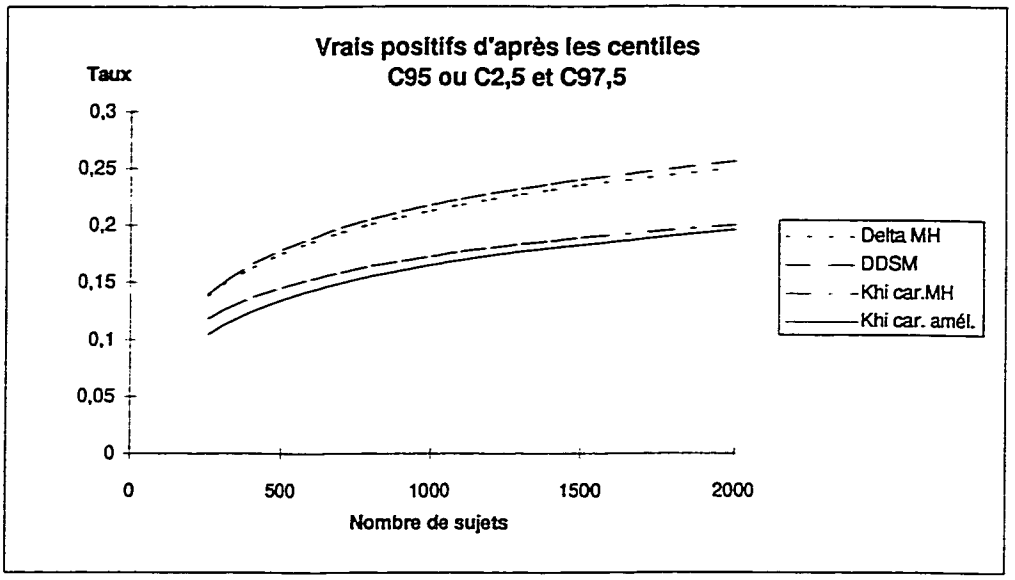


Figure 14. Taux moyens prédits de vrais positifs pour les centiles C95 ou C2,5 et C97,5.

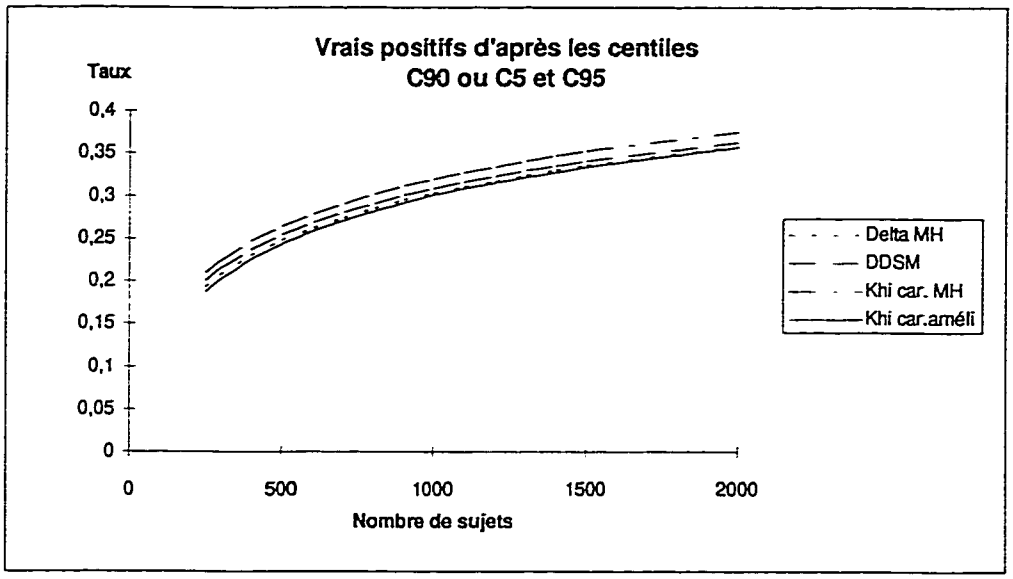


Figure 15. Taux moyens prédits de vrais positifs pour les centiles C90 ou C5 et C95.

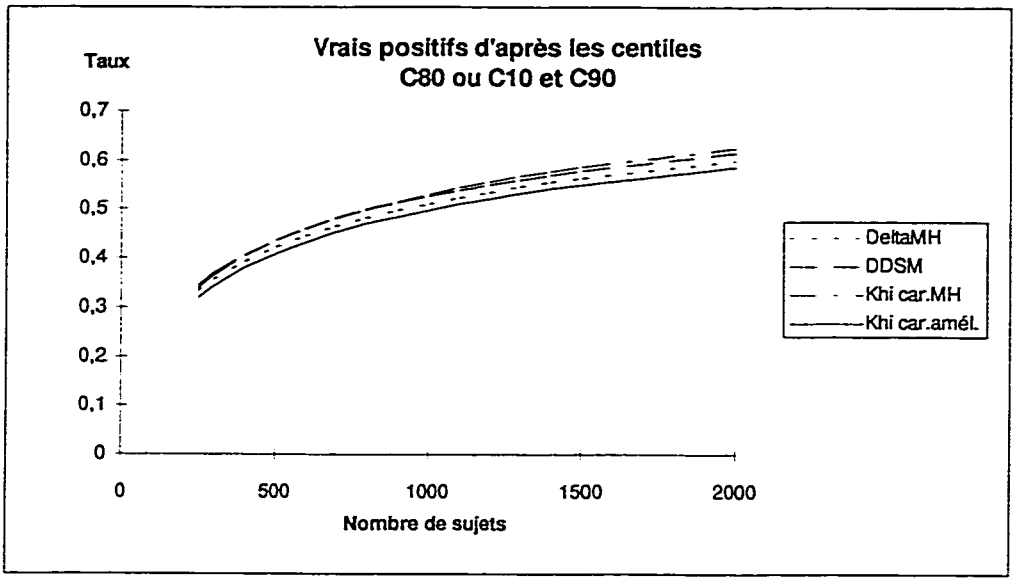


Figure 16. Taux moyens prédits de vrais positifs pour les centiles C80 ou C10et C90.

Taux objectifs basés sur des critères fixés a priori

En ce qui concerne les critères fixés a priori, l'examen du tableau 19d et de la figure 13 montre un renversement de situation par rapport aux taux relatifs correspondants. Pour les échantillons de plus de 600 sujets, les deux méthodes les plus aptes à détecter de vrais positifs sont la méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique. La méthode du delta de Mantel-Haenszel et la méthode de standardisation modifiée s'avèrent les plus efficaces pour les échantillons de moins de 600 sujets. Le renversement de situation tient au fait que le point de référence utilisé pour le calcul des taux est différent et que le nombre d'items détectés demeure à peu près constant. Le phénomène a été illustré au tableau 17 (p. 157).

La plus grande efficacité de la méthode du khi carré de Mantel-Haenszel et de la méthode de régression logistique pour les échantillons de plus de 600 sujets s'accompagne d'un taux considérable de faux positifs. De même, la plus grande efficacité de la méthode du delta de Mantel-Haenszel et de la méthode de standardisation modifiée pour les échantillons de moins de 600 sujets s'accompagne d'un taux de faux positifs très élevé. Pour les échantillons de plus de 600 sujets, leur capacité de détecter des vrais positifs est un peu moins grande et il y a une diminution importante du taux de faux positifs. Eu égard au fait que le taux de vrais positifs varie peu pour les échantillons de plus de 600 sujets, que le taux de faux positifs est nettement moins élevé que le taux observé pour la méthode du khi carré de Mantel-Haenszel ou la méthode de régression logistique et qu'il varie peu pour les échantillons de plus de 1000, ces méthodes peuvent s'avérer intéressantes pour qui veut éviter une trop grande proportion d'items identifiés à tort comme ayant un fonctionnement différentiel. Inversement, la méthode du khi carré de

Mantel-Haenszel et la méthode de régression logistique peuvent s'avérer intéressantes pour les échantillons de moins de 600 sujets, en raison du taux de faux positifs plus petit.

Pour ce qui est de la grandeur des taux de vrais positifs en fonction de la taille des échantillons (tableau 19d, p. 167), ils passent de 44,2 à 31,5 pour cent pour la méthode du delta de Mantel-Haenszel et les taux de faux positifs, de 24,8 à 1,5 pour cent. Dans le cas de la méthode de standardisation modifiée, les taux de vrais positifs passent de 44,6 à 31,7 pour cent et les taux de faux positifs, de 22,9 à 1,2 pour cent. Autrement dit, la méthode de standardisation modifiée présente des taux de vrais positifs et des taux de faux positifs voisins des taux observés pour la méthode du delta de Mantel-Haenszel. En ce qui concerne la méthode du khi carré de Mantel-Haenszel, les taux de vrais positifs vont de 17,5 à 86,0 pour cent et les taux de faux positifs, de 4,9 à 20,0 pour cent. Pour la méthode de régression logistique, les taux de vrais positifs passent de 20,5 à 84,1 pour cent et les taux de faux positifs correspondants, de 7,9 à 25,2 pour cent. Ici encore, les taux de vrais positifs sont plus faibles que nous l'aurions cru. Leur faiblesse s'explique essentiellement par la faiblesse de l'écart qui sépare les différences de fonctionnement des items qui ont un fonctionnement différentiel non négligeable de ceux qui ont un fonctionnement différentiel négligeable.

Comme dans le cas des centiles, nous avons procédé à une analyse de tendance, à l'aide du modèle logarithmique. Les équations de régression obtenues pour chaque méthode sont indiquées au tableau 21. La figure 17 reproduit les courbes théoriques qui décrivent la relation entre le taux de vrais positifs et la grandeur des échantillons. Les taux prédits de vrais positifs sont fournis à l'annexe B (voir le tableau 9d).

Tableau 21. Équations de régression logarithmique des taux de vrais positifs en fonction de la grandeur des échantillons pour des critères fixés a priori.

Indice	Critère	Équation	Coefficient R^2
Δ_{MH}	**1,00 ou -1,00**	y = 0,3338ln(x) - 1,7032	0,9852
DDSM	**0,075 ou -0,075**	y = 0,3088ln(x) - 1,5311	0,9860
X^2_{MH}	**0,3841**	y = -0,0599ln(x) + 0,7703	0,9965
$X^2_{amél.}$	**0.5991**	y = -0,0615ln(x) + 0,7857	0,9993

Une fois de plus, les courbes théoriques de régression font ressortir la similitude entre la méthode du delta de Mantel-Haenszel et la méthode de standardisation modifiée. La différence entre les taux prédits et les taux observés est inférieure à un demi pour cent pour la méthode du delta de Mantel-Haenszel et la méthode de standardisation modifiée. Elle varie de 2,5 à 3,5 pour cent selon les méthodes et la taille des groupes, pour la méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique. Les courbes théoriques indiquent également que la méthode la plus efficace change pour des échantillons de 550 sujets. Les données observées indiquaient un changement pour des échantillons d'environ 575 sujets.

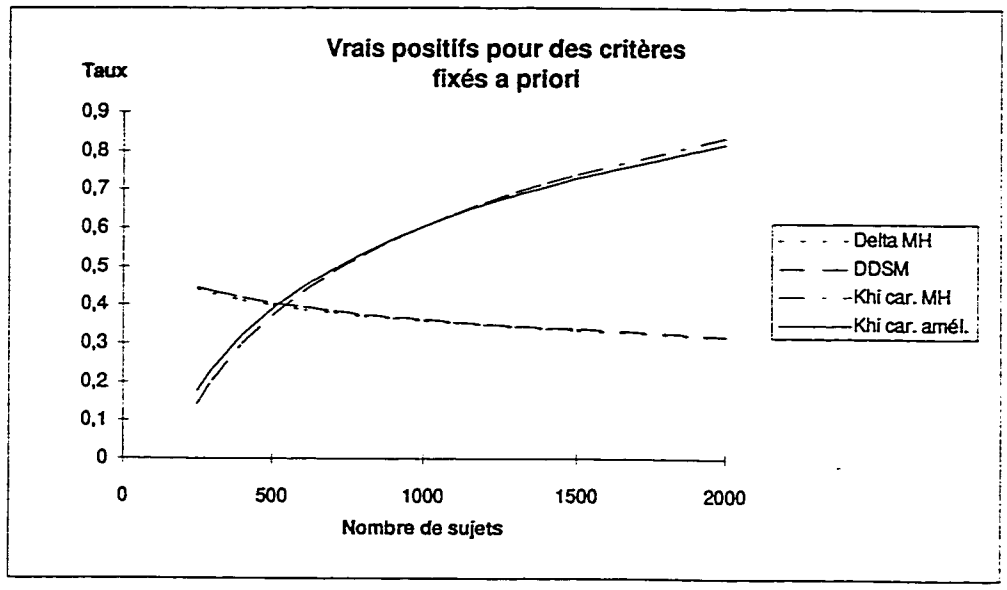


Figure 17. Taux moyens prédits de vrais positifs pour des critères fixés a priori.

Quant à savoir ce qui arriverait si nous nous étions limités à ne considérer comme ayant un fonctionnement différentiel non négligeable que les cinq items qui ont un fonctionnement différentiel assez élevé dans la population, les taux de vrais positifs seraient plus grands, quelle que soit la méthode. Les taux de faux positifs augmenteraient en conséquence. Le renversement de situation se maintiendrait. La méthode du delta de Mantel-Haenszel et la méthode de standardisation modifiée s'avéreraient plus efficaces pour détecter les vrais positifs que la méthode du khi carré de Mantel-Haenszel ou la méthode de régression logistique pour les échantillons de moins de 600 sujets. La méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique le seraient pour des échantillons de plus de 600 sujets.

CHAPITRE VI

DISCUSSION

La présente étude avait pour but d'examiner la fidélité et l'efficacité de la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991a, 1991b, 1992a, 1992b, 1993 et 1995) et de la comparer à la méthode de Mantel-Haenszel et à la méthode de régression logistique. Dans cette perspective, nous avons examiné la stabilité des indices et leur la capacité de détecter des items qui ont un fonctionnement différentiel dans la population. Fidélité et efficacité ont été examinées pour des échantillons de taille variée. Chaque échantillon se compose de deux groupes égaux de sujets. L'utilisation d'échantillons constitués de groupes inégaux devrait donner des résultats différents. L'ampleur des différences dépendra de l'importance de la disproportion. Plus celle-ci est importante, plus les différences risquent de s'avérer importantes. Les résultats devront donc être évalués en gardant à l'esprit que les groupes comparés comptent un nombre égal de sujets. Dans les pages qui suivent, nous discuterons des résultats observés. Nous aborderons d'abord la question de la stabilité des indices, puis celle de la validité et de la fidélité des décisions, et enfin, celle de l'efficacité.

Stabilité des indices

En ce qui concerne la stabilité des indices, la méthode du delta de Mantel-Haenszel offre des indices généralement plus stables que la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995). Quelle que soit la taille des échantillons, les deux méthodes peuvent fournir des coefficients de stabilité

pratiquement aussi élevés. C'est au niveau des coefficients les plus faibles que se manifeste la différence. Ceci dit, la différence entre les coefficients moyens ou médians n'est pas significative pour des échantillons de même grandeur. Compte tenu du peu de différence entre les deux méthodes, il nous apparaît que la méthode de standardisation modifiée peut être utilisée au lieu de la méthode du delta de Mantel-Haenszel sans que la stabilité des indices ne devienne un problème plus important qu'il ne l'est avec la méthode du delta de Mantel-Haenszel.

La méthode du delta de Mantel-Haenszel et la méthode de standardisation modifiée présentent des coefficients de stabilité généralement plus élevés que la méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique. Lorsqu'on considère le nombre de fois qu'un item est détecté pour chaque item et chaque grandeur d'échantillon et qu'on utilise les centiles comme critère de décision, on constate qu'il y a peu de différence entre les méthodes. Toutes pourraient avoir de la difficulté à classer les items de la même façon, peu importe la façon dont chaque méthode classe les items. Nous en déduisons que les différences observées entre les méthodes au niveau des coefficients de stabilité ont peu de conséquence pratique. De fait, celles-ci nous paraissent imputables d'abord et avant tout au type d'indices en cause, orientés ou non orientés selon le cas. L'utilisation d'indices orientés (Δ_{MH} ou DDSM) réduit la variabilité possible dans le classement des items. À cela s'ajoute l'utilisation de coefficients de corrélation de rangs pour les indices non orientés (X_{MH} et $X_{amél.}$) et l'utilisation de coefficients de corrélation de Pearson pour les indices orientés (Δ_{MH} ou DDSM). Toutefois, la simple utilisation de coefficients de corrélation de rangs pour les deux types d'indices n'éliminerait pas toutes les différences.

La présente étude montre également une relation entre la grandeur et l'étendue des coefficients de stabilité et le nombre de sujets dans les échantillons. L'existence d'une relation entre la fidélité des indices et la grandeur des échantillons n'a rien de nouveau. Perlman, Bezruezko, Junker, Reynolds, Rice et Schulz (1988) ont déjà observé une telle relation avec la méthode de Mantel-Haenszel. Nous avons montré qu'elle s'applique aussi à la méthode de standardisation modifiée, à la méthode du khi carré de Mantel-Haenszel et à la méthode de régression logistique. Nous espérions que la méthode de standardisation modifiée aurait été moins touchée par la diminution des effectifs. En fait, elle n'est ni meilleure ni pire que les autres. La recherche de moyens autres que l'augmentation du nombre de sujets pour accroître la stabilité des indices et la constance des décisions, et partant, la généralisabité des études de fonctionnement différentiel demeure donc ouverte.

La méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995) utilise un procédé de lissage qui repose sur la moyenne locale. Pour la présente étude, nous avons fixé à 61 le nombre de points d'estimation des probabilités de réponses et nous avons laissé au programme le soin de fixer la grandeur de l'intervalle pour le procédé de lissage. Le nombre de points a été tenu constant pour toutes les grandeurs d'échantillon. La grandeur de l'intervalle variait en fonction de la grandeur des échantillons. Dans l'état actuel de la recherche, nous ignorons la grandeur de l'intervalle et le nombre optimal de points requis pour atténuer au maximum l'effet des cas extrêmes sur les indices de fonctionnement différentiel. À plus forte raison, nous ignorons leur effet en relation avec la taille des échantillons et le nombre d'items contenu dans le test, deux variables susceptibles d'avoir un effet sur les indices de fonctionnement différentiel. Un meilleur usage de

ces deux composantes pourrait peut-être améliorer la fidélité de la méthode de standardisation modifiée. Des recherches seraient nécessaires pour clarifier la question. La prise en considération de l'erreur due aux fluctuations d'échantillonnage, au moment d'utiliser les indices de fonctionnement différentiel, pourrait également s'avérer efficace. Le procédé vaut également pour la méthode du delta de Mantel-Haenszel. L'hypothèse est à vérifier dans les deux cas.

Validité et fidélité des décisions

L'examen de la fréquence de détection de chaque item pour chaque méthode et chaque grandeur d'échantillon confirme l'existence d'une relation entre la détection des items qui ont un fonctionnement différentiel et le nombre de sujets dans les échantillons. Ici encore, l'existence d'une telle relation n'a rien de nouveau. Pang, Tian et Boss (1994) ont fait état de fréquences de détection qui augmentent avec la taille des échantillons pour la méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique. Ce qui est nouveau, c'est la mise en évidence d'une relation entre la fréquence de détection des items, la grandeur des différences dans le fonctionnement des items et le critère de décision utilisé. À cet égard, nous constatons des divergences selon qu'on utilise des centiles ou des critères fixés a priori. Dans le cas des centiles, la fréquence de détection des items tend à croître avec l'augmentation du nombre de sujets dans les échantillons ou à se maintenir au même niveau pour tous les items qui accusent des différences de fonctionnement non négligeables dans la population. Il y a augmentation pour les items qui accusent les différences de fonctionnement les plus grandes et maintien pour les items avec les différences de fonctionnement moins grandes. Par contre, la fréquence de détection tend à décroître avec l'augmentation de la taille des échantillons lorsque les items présentent des différences de fonctionnement négligeables dans la population. Cette tendance est particulièrement évidente pour les items dont les différences de fonctionnement ne dépassent pas les différences observées entre deux groupes issus d'une même sous-population. Le phénomène existe pour toutes les méthodes à l'étude lorsqu'on utilise des centiles comme critère de décision. Lorsqu'on utilise un critère fixé a priori, l'unité pour le delta de Mantel-Haenszel, la valeur de 0,075 pour la méthode de standardisation modifiée, la valeur critique au seuil de 0,05 pour la

méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique, on observe un phénomène similaire pour la méthode du delta de Mantel-Haenszel et la méthode standardisation modifiée. Dans le cas de la méthode du khi carré de Mantel-Haenszel et de la méthode de régression logistique, la fréquence de détection des items tend à croître ou à se maintenir au même niveau pour tous les items, même ceux qui n'ont pas été retenus comme ayant un fonctionnement différentiel dans la population. Toutefois, la fréquence de détection des items qui ont un fonctionnement négligeable dans la population croît moins qu'elle ne croît pour les items retenus comme ayant un fonctionnement différentiel non négligeable. Ces résultats donnent à penser qu'il pourrait y avoir avantage à examiner la possibilité d'optimiser l'efficacité des méthodes à partir du critère de décision et plus particulièrement, d'un critère qui varie en fonction de la taille des échantillons.

En ce qui concerne la validité des décisions pour l'ensemble des items qui composent le test, elle a été évaluée à partir de la relation entre les items détectés dans les échantillons et les items identifiés comme ayant un fonctionnement différentiel non négligeable dans la population. De plus, nous avons considéré le nombre de décisions correctes. Dans les deux cas, nous constatons une relation entre la grandeur des échantillons et la validité des décisions. Nous constatons également qu'il y a peu de différences entre les méthodes sous ce rapport. Tout au plus note-on une propension à obtenir plus de décisions valides lorsqu'on utilise la méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique avec des échantillons de 250 sujets et des critères fixes au lieu de la méthode du delta de Mantel-Haenszel ou de la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991a, 1991b, 1992a, 1992b, 1993 et 1995). Cette particularité apparaît uniquement lorsqu'on se sert de critère fixes.

Il est par ailleurs difficile de comparer les résultats observés à ceux d'autres recherches. Peu de recherche, s'il en est, ont évalué la validité des méthodes à l'étude sous cet angle.

L'un des buts de l'étude était de comparer la méthode de standardisation modifiée à la méthode de Mantel-Haenszel et à la méthode de régression logistique. Nous ne pouvons que constater qu'elle offre des décisions généralement aussi valides que les autres méthodes à l'étude et qu'elle se comporte généralement comme la méthode du delta de Mantel-Haenszel. L'autre but de l'étude était de déterminer la grandeur d'échantillon souhaitable pour s'assurer de décisions fiables et valides. L'augmentation des coefficients de stabilité et de validité en fonction de la taille des échantillons, l'augmentation de la fréquence de détection de chaque item et du nombre de décisions correctes et la diminution de la variabilité des coefficients et des fréquences confirment l'utilité d'échantillons relativement grands pour s'assurer que les analyses de fonctionnement différentiel auront une généralisabilité minimalement acceptable. Sous ce rapport, il devient évident que des échantillons de 250 sujets sont nettement insuffisants pour s'assurer de décisions fiables et valides. Des échantillons de 500 sujets demeurent problématiques et des échantillons de 1000 à 2000 sujets s'avèrent nettement préférables. Ceci dit, l'augmentation de nombre de sujets dans les échantillons se traduit généralement par une augmentation des vrais positifs et des vrais négatifs. Pour cette raison, la détermination du nombre souhaitable de sujets devrait reposer en dernière analyse sur l'efficacité relative des méthodes.

Efficacité relative des méthodes

Nous avons examiné l'efficacité relative des méthodes à partir de taux de détection et nous avons fait état de courbes théoriques qui estiment la relation entre les taux d'identification correcte ou les taux de vrais positifs et la grandeur des échantillons. Les taux sont tributaires des critères utilisés pour déterminer quels items ont un fonctionnement différentiel dans les échantillons. Les courbes théoriques présupposent le choix d'un modèle de régression. On peut s'interroger sur le choix du modèle et des critères de décision.

En ce qui concerne le modèle de régression, nous avons retenu le modèle logarithmique, parce que c'est celui qui s'ajustait le mieux à la totalité des données. Ce modèle présuppose une augmentation continue de la variable prédite. Cette variable est constituée par les taux de détection. Ces derniers ont une valeur maximale : l'unité ou 100 pour cent selon l'échelle utilisée. Dans ces conditions, les modèles log-linéaires ou logit, le modèle de régression logistique, un modèle de lissage ou tout autre modèle capable de limiter l'étendue de la variable prédite conviendraient mieux au plan théorique. Pour choisir un autre modèle ou maximiser la précision des taux prédits à l'extrémité de l'échelle, il aurait fallu disposer d'échantillons de plus de 2000 sujets. Toutefois, en autant qu'on s'en tient à des échantillons inférieurs à 2000 sujets l'utilisation du modèle logarithmique ne pose pas de problèmes majeurs. L'écart entre les valeurs prédites et les valeurs observées est toujours inférieur à 3,4 pour cent pour les taux d'identification correcte ou les taux de vrais positifs. Pour ces derniers, il est généralement inférieur à 1,5 pour cent.

Quant au choix des critères de décision utilisés pour déterminer quels items ont un fonctionnement différentiel, les critères a priori reposent sur la pratique en usage pour la méthode de Mantel-Haenszel et la méthode de régression logistique. Pour la méthode de standardisation modifiée, il nous a fallu établir un critère à l'avenant. D'autres études seraient nécessaires pour voir jusqu'à quel point celui-ci peut convenir en toutes circonstances. Pour ce qui est des centiles, ces derniers ne reflètent pas la pratique en usage. En général, on utilise les valeurs critiques au seuil de 0,01 ou de 0,05 pour les indices basés sur des tests statistiques et, à défaut de valeurs fixées à priori, les valeurs d'indices à 1,96 écart-type pour les indices basés sur une mesure de l'ampleur des différences dans le fonctionnement des items. Nous avons postulé que les méthodes à l'étude étaient en mesure de classer les items de façon assez semblable dans les échantillons et dans la population. Leur incapacité de le refléter dans les taux de détection tenait à l'utilisation de critères qui ne tiennent pas assez compte du classement des items dans une distribution d'indices. L'utilisation de centiles permettait l'identification d'un nombre constant d'items d'un échantillon à l'autre. Nous pensions que cela faciliterait la comparaison. L'utilisation des valeurs d'indices à 1,96 écart-type est de nature à augmenter le taux de vrais positifs et à diminuer celui des faux positifs, mais elle risque de donner lieu à plus de variabilité dans le nombre d'items détectés.

Nous avons examiné la capacité de détecter des items qui ont un fonctionnement différentiel sous deux aspects : le taux moyen d'identification correcte et le taux moyen de vrais positifs pour les 100 échantillons de même taille. Pour ce qui est des taux basés sur des centiles, les deux approches pointent dans la même direction. La méthode de standardisation modifiée s'avère généralement aussi efficace que la méthode du delta de Mantel-Haenszel sinon plus. Toutefois,

c'est la méthode du khi carré de Mantel-Haenszel qui s'avère la plus efficace, abstraction faite des taux établis après application des centiles C_{95} ou $C_{2,5}$ et $C_{97,5}$

Qu'il s'agisse du taux d'identification correcte ou du taux de vrais positifs, il y a peu de différence entre les méthodes pour une même grandeur d'échantillon. Nous nous attendions à plus d'écart. Le peu de différence s'explique par l'étroite relation qui existe entre les méthodes au plan conceptuel. Pour les taux de vrais positifs, l'écart maximal entre la méthode la plus efficace et la méthode la moins efficace varie entre 2,210 et 6,040 pour cent selon les centiles utilisés comme critère de décision. Une telle différence a peu de conséquences pratiques, puisqu'elle ne donne pas l'assurance de détecter un item de plus avec le même nombre de sujets. Pour avoir une telle assurance, il faudrait que la différence de vrais positifs s'élève à 6,667 pour cent. Dans le même ordre d'idée, il faudrait que les taux d'identification correcte accusent une différence qui varie entre 8,333 et 25 ou 33 pour cent selon les centiles utilisés comme critère de décision pour que les différences observées se traduisent par la détection d'un item de plus.

Pour ce qui est de l'efficacité des méthodes pour des critères fixés a priori, la situation est plus complexe. La méthode de standardisation modifiée est aussi efficace que la méthode du delta de Mantel-Haenszel, mais les deux méthodes se comportent de façon diamétralement opposée à la méthode du khi carré de Mantel-Haenszel ou à la méthode de régression logistique. La méthode du delta de Mantel-Haenszel et la méthode de standardisation modifiée identifient un nombre d'items d'autant plus grand qu'il y a moins de sujets dans les échantillons. Le nombre total d'items détectés diminue avec l'augmentation du nombre de sujets dans les échantillons. Il en résulte une légère diminution des vrais positifs et, toutes proportions gardées, beaucoup moins de

faux positifs. La méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique identifient comme ayant un fonctionnement différentiel un nombre d'items d'autant plus grand que les échantillons comptent plus de sujets. Il y a plus de vrais positifs, mais aussi plus de faux positifs. Le fait que les méthodes basées sur une mesure de l'ampleur des différences dans le fonctionnement des items et les méthodes basées sur un test statistique ont un comportement divergent plaide en faveur d'une combinaison des deux types de méthodes avec de petits échantillons. À défaut de combiner les deux types de méthodes, le choix d'une méthode de préférence à une autre dépendra de l'importance accordée à la détection de vrais positifs et du taux de faux positifs que l'on est prêt à tolérer.

Nonobstant le critère de décision utilisé, il appert que des échantillons de 500 sujets constituent la grandeur minimale nécessaire pour s'assurer de détecter une proportion minimale acceptable d'items avec un fonctionnement différentiel dans la population. Des échantillons de 1000 sujets seraient nettement préférables. Avec des échantillons de 500 sujets, la proportion d'items détectés qui ont un fonctionnement différentiel dans la population, la proportion d'identification correcte donc, est de trois sur cinq par rapport à la totalité des items détectés (25 pour cent de vrais positifs) pour les centiles C_{90} ou C_5 et C_{95} et de un sur deux (40 pour cent de vrais positifs) pour les centiles C_{80} ou C_{10} et C_{90}. Avec des échantillons de 1000 sujets, elle passe à trois sur quatre pour le premier critère de décision (30 pour cent de vrais positifs) et à deux sur trois pour le deuxième critère (50 pour cent de vrais positifs). Par ailleurs, des échantillons de 500 ou de 1000 sujets permettent de maximiser la proportion de vrais positifs tout en minimisant la proportion de faux positifs lorsqu'on utilise des critères fixés a priori.

CHAPITRE V11

CONCLUSION

La présente étude a vérifié la fidélité et l'efficacité de méthodes d'analyse du fonctionnement différentiel des items applicables à des échantillons de taille réduite. Les méthodes considérées sont la méthode de Mantel-Haenszel, la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995) et la méthode de régression logistique. Les indices retenus sont le khi carré et le delta de Mantel-Haenszel, la différence de difficulté standardisée modifiée et le khi carré d'amélioration pour la méthode de régression logistique. Fidélité et efficacité ont été étudiées pour des échantillons de 250, de 500, de 1000 et de 2000 sujets. Chaque échantillon se compose d'un nombre égal de personnes dans chacun des deux groupes comparés.

L'étude repose sur des données réelles. Le test analysé a déjà fait l'objet d'études du fonctionnement différentiel des items. Les items qui affichaient une différence importante de fonctionnement ont été éliminés et les autres ont été agencés de manière à en neutraliser les effets. Par conséquent, les items qui présentent des différences de fonctionnement non négligeables dans la population ont en réalité des différences relativement faibles et l'écart qui les sépare des items dont les différences de fonctionnement sont négligeables n'est pas très grand. Ces conditions sont de nature à atténuer la stabilité des indices et la constance des décisions d'un échantillon à un autre de même taille. Elles sont également de nature à rendre plus difficile la détection des items qui ont un fonctionnement différentiel non négligeable dans la population. Par contre, elles témoignent de la valeur des méthodes, celles-ci étant en mesure

de fonctionner de façon satisfaisante dans des conditions difficiles. De plus, elles montrent que la méthode de standardisation modifiée constitue un substitut valable pour remplacer la méthode du delta ou du khi carré de Mantel-Haenszel et la méthode de régression logistique. Les pages qui suivent résument les conclusions de l'étude. Elles en soulignent également les limites et suggèrent de nouvelles recherches.

Résumé des résultats et conclusions

La méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995) est nouvelle. La méthode de Mantel-Haenszel et la méthode de régression logistique sont reconnues à l'heure actuelle comme les plus appropriées avec de petits échantillons. La présente étude montre que la méthode de standardisation modifiée se comporte de façon analogue à la méthode du delta de Mantel-Haenszel. Il en est ainsi tant au niveau de la stabilité des indices qu'au niveau de la capacité de détecter des items qui ont un fonctionnement différentiel non négligeable dans la population.

Ces résultats suffisent pour conclure à la validité de l'approche comme moyen d'étudier le fonctionnement différentiel des items. La méthode offre des indices généralement un peu moins stables que la méthode du delta de Mantel-Haenszel. Malgré cela, sa capacité de détecter des items qui ont un fonctionnement différentiel se compare avantageusement à la méthode du delta de Mantel-Haenszel. Elle peut également se comparer avantageusement à la méthode du khi carré de Mantel-Haenszel et à la méthode de régression logistique sous certaines conditions que nous préciserons plus loin.

Quelle que soit la méthode, il y a une relation entre la grandeur des échantillons et la stabilité des indices, la constance des décisions et leur validité. Il y a également une relation entre la grandeur des échantillons et l'efficacité des méthodes. Cette dernière est non linéaire, que l'on considère le taux d'identification correcte ou le taux de vrais positifs.

En ce qui concerne l'efficacité relative des méthodes, les quatre méthodes à l'étude se comportent de façon analogue lorsqu'on utilise la valeur d'indice à des centiles prédéterminés pour déterminer quels items ont un fonctionnement différentiel. Lorsqu'on se sert de critères fixés a priori, la méthode du delta de Mantel-Haenszel et la méthode de standardisation modifiée s'avèrent à toutes fins utiles aussi efficaces l'une que l'autre, mais elles affichent un comportement à l'opposé de la méthode du khi carré de Mantel-Haenszel et de la méthode de régression logistique. Cette divergence tient à la nature des méthodes. Les méthodes basées sur l'ampleur des différences dans le fonctionnement des items voient leurs indices exagérément gonflés avec des échantillons de moins de 600 sujets. Il en résulte un taux de vrais positifs à peu près constant et un taux de faux positifs plus grand avec les échantillons plus petits et qui va décroissant avec l'augmentation du nombre de sujets dans les échantillons. Les méthodes basées sur un test statistique ont tendance à identifier un nombre plus grand d'items avec des différences minimes lorsque la taille des échantillons est plus grande. Il en résulte un taux de vrais positifs plus grand, mais aussi un taux de faux positifs nettement plus grand au fur et à mesure que croît la taille des échantillons.

Qu'il s'agisse de taux basés sur des critères fixes ou de taux basés sur des centiles, la faiblesse des taux de vrais positifs s'explique par le fait que les différences de fonctionnement que l'on

espère détecter dans la population sont petites. La faiblesse des taux de vrais positifs basés sur des centiles s'explique aussi par le choix des centiles seuils. Aucun ne permet de détecter la totalité des items qui ont un fonctionnement différentiel non négligeable. Au mieux, on ne peut détecter que 80 pour cent des items identifiés comme ayant un fonctionnement différentiel non négligeable dans la population. Au pire, on ne peut en détecter que 20 pour cent.

En ce qui concerne le nombre minimum de sujets requis dans les échantillons pour s'assurer d'indices stables et efficaces et de décisions fiables et valides, il appert que des échantillons de 250 sujets (125 sujets par groupe) sont nettement insuffisants. Des échantillons de 500 sujets (250 sujets par groupe) offrent une stabilité problématique. Des échantillons de 1000 sujets (500 sujets par groupe) représentent un minimum et des échantillons de 2000 sujets (1000 sujets par groupe) s'avèrent manifestement préférables sous ce rapport.

Limites de l'étude

La présente étude repose sur des données réelles. Par conséquent, la généralisabilité des résultats se limite à des tests composés d'items qui présentent des caractéristiques semblables à celles des items du test analysé. Il en est ainsi tant pour les items qui ont un fonctionnement différentiel non négligeable que pour ceux qui ont un fonctionnement différentiel négligeable. Ces caractéristiques comprennent la difficulté et la discrimination des items, mais aussi l'ampleur des différences dans le fonctionnement des items et le type de fonctionnement différentiel observé (fonctionnement différentiel uniforme ou non uniforme).

L'efficacité des méthodes, telle que déterminée par les taux de détection, dépend de l'ampleur des différences jugées non négligeables dans la population. Leur petitesse a un effet certain sur les taux observés. L'écart entre les items qui présentent un fonctionnement différentiel non négligeable et ceux qui présentent des différences de fonctionnement négligeable a également un effet. Cet effet se manifeste tant au niveau de la stabilité des indices qu'au niveau de l'efficacité. Un écart important facilite la détection d'items qui ont effectivement un fonctionnement différentiel. S'il est faible, il la rend plus difficile. L'efficacité des méthodes est aussi tributaire des critères retenus pour déterminer quels items ont un fonctionnement différentiel dans les échantillons.

En ce qui concerne plus particulièrement la méthode de standardisation modifiée, les résultats observés dépendent aussi des conditions dans lesquelles la méthode a été appliquée. À cet égard, le nombre de points d'estimation des probabilités de réponses en fonction de l'habileté des sujets,

la grandeur de l'intervalle pour l'application du procédé de lissage et la prise en considération des abstentions comme une sixième réponse possible, en plus des cinq réponses suggérées, sont susceptibles d'avoir eu un effet. Il en va de même pour le choix du critère fixe de décision utilisé pour la présente étude.

Enfin, l'étude repose sur un ensemble de postulats qui, s'ils ne se réalisent pas, compromettent la validité et la généralisabilité des résultats. À cet égard, il est postulé que le score total constitue une mesure valable et non biaisée de l'habileté des sujets. Il est également postulé que les items sont unidimensionnels ou qu'ils fonctionnent comme des items unidimensionnels et que les réponses à un item sont indépendantes des réponses aux autres items. Enfin, il est postulé que le trait latent mesuré est continu et que la perte d'information qui résulte d'une notation dichotomique est négligeable.

Par ailleurs, en ce qui concerne les courbes théoriques de régression logarithmique décrivant la relation entre les taux d'identification correcte ou les taux de vrais positifs et le nombre de sujets dans les échantillons, il ne fait pas de doute qu'une plus grande diversité dans la taille des échantillons inférieurs à 2000 sujets et l'utilisation d'échantillons de plus de 2000 sujets auraient pu influencer le choix du modèle de régression et surtout la précision des valeurs prédites.

Malgré ses limites, la présente étude contribue à l'avancement des connaissances en ce qui a trait aux méthodes d'analyses du fonctionnement différentiel des items. Les items identifiés comme ayant un fonctionnement différentiel non négligeable dans la population présentent des différences petites mais valides. Dans ces conditions plutôt difficiles, elle montre que la méthode

de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991a, 1991b, 1992a, 1992b, 1993 et 1995) peut fournir des indices dont la stabilité et l'efficacité se comparent à celles des autres méthodes étudiées, et plus particulièrement, à la méthode du delta de Mantel-Haenszel. Or, le modèle de Ramsay joint à la méthode de standardisation présente plusieurs avantages. Nous les avons énumérés au chapitre III (p. 45). Mentionnons entre autres la prise en compte des mauvaises réponses pour estimer l'habileté des sujets et la possibilité d'étudier le fonctionnement différentiel des leurres conjointement à l'étude du fonctionnement différentiel des items. La prise en compte des mauvaises réponses devrait accroître la précision de l'estimation de l'habileté des sujets les plus faibles. L'étude du fonctionnement différentiel des items à l'aide de la méthode de standardisation modifiée et de Testgraf, le programme mis au point par Ramsay, fournit des indices de fonctionnement différentiel pour les bonnes réponses, mais aussi pour les leurres. L'examen des indices de fonctionnement différentiel des leurres facilite la formulation d'hypothèses sur les causes du fonctionnement différentiel et les moyens d'y remédier. Enfin, elle met en évidence l'existence d'un lien entre le critère de décision, l'ampleur des différences de fonctionnement que l'on espère détecter dans la population et l'efficacité des méthodes. Ce faisant, elle montre que la taille des échantillons peut-être fixée en fonction de ces facteurs.

Suggestions de recherches

Parce que l'étude repose sur des données réelles, que le test a été conçu de manière à équilibrer les différences de fonctionnement des items en faveur de chaque groupe et que celles-ci sont relativement faibles, il y aurait lieu de la reprendre avec d'autres tests dont la longueur et les caractéristiques diffèrent. Idéalement, celles-ci devraient différer tant pour ce qui est de la difficulté et de la discrimination des items que pour l'ampleur des différences dans le fonctionnement des items et l'écart qui sépare les items qui ont un fonctionnement différentiel non négligeable de ceux qui ont un fonctionnement différentiel négligeable. Il y aurait également lieu de répéter l'étude avec plus de grandeurs d'échantillon de moins de 2000 sujets afin d'accroître la précision des valeurs prédites dans les taux de détection pour les grandeurs d'échantillon entre 1000 et 2000 sujets.

La présente étude a examiné l'efficacité des méthodes en fonction de centiles prédéterminés comme critère décisionnel. L'utilisation de valeurs d'écart-type est plus usuelle. Pour cette raison, nous estimons utile de reprendre l'étude avec des valeurs d'indice à 1,96 écart-type comme critère de décision. Cette étude pourrait se limiter à la méthode du delta de Mantel-Haenszel et la méthode de standardisation modifiée par l'application du modèle de Ramsay et mettre les résultats en relation avec la méthode du khi carré de Mantel-Haenszel et la méthode de régression logistique lorsqu'on utilise la valeur critique au seuil de 0,05 ou de 0,01. Enfin, il y aurait lieu d'étudier l'efficacité des méthodes en fonction du type de fonctionnement différentiel qui caractérise les items (fonctionnement différentiel uniforme ou non uniforme).

La méthode de standardisation modifiée nécessiterait aussi des études plus poussées pour déterminer l'effet du nombre de points d'estimation des probabilités de réponses et déterminer le nombre requis pour maximiser son efficacité eu égard à la longueur des tests et à la grandeur des échantillons. L'effet de la grandeur de l'intervalle pour l'application du procédé de lissage mériterait aussi un examen. Toutefois, ce point concerne moins l'étude du fonctionnement différentiel des items que le modèle de Ramsay. Pour revenir à l'étude du fonctionnement différentiel des items, le choix du critère fixe de décision utilisé pour la présente étude nécessiterait d'être étudié à nouveau. Il y aurait également lieu d'examiner la possibilité de faire varier le critère fixe de décision en fonction de la taille des échantillons, ceci afin de maximiser la détection des vrais positifs tout en minimisant celle des faux positifs.

Enfin, la difficulté et la discrimination des items, la différence d'habileté des groupes comparés et leur plus ou moins grande disproportion sont susceptibles d'avoir un effet sur la méthode de standardisation modifiée comme elles en ont sur la méthode de Mantel-Haenszel et la méthode de régression logistique. Pour cette raison, nous en recommandons l'étude, cette fois, avec des données simulées.

Une dernière recommandation s'impose. La présente étude avait pour but de comparer la méthode de standardisation modifiée par l'application du modèle de Ramsay (1989, 1991 a, 1991 b, 1992 a, 1992 b, 1993 et 1995) à la méthode de Mantel-Haenszel et à la méthode de régression logistique. Cette méthode a été suggérée par Dorans, Schmitt et Bleistein (1992, p. 311) pour remplacer la méthode de standardisation lorsque les échantillons sont de taille réduite. Cette dernière est considérée comme inadéquate dans de telles conditions, parce qu'elle

fournit des indices exagérément gonflés. Nous avons observé une efficacité comparable à la méthode du delta de Mantel-Haenszel après avoir modifié le critère de décision. Liu, Dorans et Ramsay (1995) ont pu constater que la méthode de standardisation modifiée fournissait des indices qui varient peu de la méthode de standardisation régulière. Nous avons nous-même pu constater que la méthode de standardisation modifiée classait un nombre exagéré d'items comme ayant un fonctionnement différentiel lorsque nous utilisions la valeur de 0,05 comme critère de décision. Dans les circonstances, il y aurait lieu de voir si la méthode de standardisation régulière ne conviendrait pas en modifiant le critère de décision. Il y aurait également lieu de comparer la méthode de standardisation modifiée à la méthode de standardisation régulière pour voir si elle lui est supérieure. Bref, il pourrait s'avérer utile de reprendre la présente étude en y incluant la méthode de standardisation régulière.

RÉFÉRENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, impact bias, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29(1), 67-91.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. Dans P. W. Holland et H. Wainer (Eds.), Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Angoff, W. H. (1988). Proposals for theoritical and applied development in measurement. Applied Measurement in Education, 1(3), 215-222.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. Dans R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Angoff, W. H., et Ford, S. E. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10(2), 95-106.

Baghi, H., et Ferrara, S. F. (1989). A comparison of IRT, delta plot, and Mantel-Haenszel techniques for detecting differential item functioning across subpopulations in the Maryland Test of Citizenship Skills. Texte présenté à la réunion annuelle de l'American Educational Research Association, San Francisco, CA. ED 324 364.

Baghi, H., et Ferrara, S. F. (1990). Detecting differential item functioning using IRT and Mantel-Haenszel techniques: Implementing procedures and comparing results. Texte présenté à la conférence annuelle de l'Eastern Educational Research Association, Clearwater, FL. ED 325 479.

Baker, F. B. A. (1981). A criticism of Scheuneman's item bias technique. Journal of Educational Measurement, 18, 59-62.

Berk, R. A. (Ed.). (1982). Handbook of methods for detecting test bias, Baltimore, MD: The Johns Hopkins University Press.

Brown, P. (1992). An empirical study of the consistency of differential item functioning detection. (Thèse de doctorat inédite. University of Ottawa).

Camilli, G. (1979). A critique of the chi-square method for assessing item bias. Texte inédit. Laboratory of Educational Research, University of Colorado. Boulder, CO.

Camilli, G., et Shepard, L. A. (1987). The inadequacy of ANOVA for detecting test bias. Journal of Educational Statistics, 12(1), 87-99.

Camilli, G., et Smith, J. K. (1990). Comparison on the Mantel-Haenszel test with a randomized and a Jackknife test for detecting biased items. Journal of Educational Statistics, 15(1), 53-67.

Cardall, C., et Coffman, W. E. (1964). A method for comparing the performance of different groups on the items in a test. (College Entrance Examination Board Research and Development Report 64-65, No. 9; ETS Research Bulletin 64-61). Princeton, NJ: Educational Testing Service.

Clauser, B. E., Mazor, K. M., et Hambleton, R. K. (1991 b). Examination of various influences on the Mantel-Haenszel statistics. Texte présenté à la réunion annuelle de l'American Educational Research Association, Chicago, IL. ED 331 876.

Clauser, B. E., Mazor, K. M., et Hambleton, R. K. (1991 c). Influence of the criterion variable on the identification of differentially functioning test items using the Mantel-Haenszel statistic. Applied Psychological Measurement, 15(4), 353-359.

Clauser, B. E., Mazor, K. M., et Hambleton, R. K. (1991 a). The effects of score group on the Mantel-Haenszel procedure. (Laboratory of Psychometric and Evaluative Research Report No. 226). Amherst, MA: University of Massachusetts.

Cohen, J. (1988). Statistical power analysis for behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Collins, L. M., Cliff, N., McCormick, D. J., et Zatkin, J. L. (1986). Factor recovery in binary data sets: a simulation. Multivariate Behavioral Research, 21(3), 377-391.

Crocker, L., et Algina, J. (1986). Introduction to classical and modern test theory. New York, NY: Holt, Rinehart and Winston.

Dann, L., Irvine, S. H., et Collis, J. M. (Eds.). (1991). Advances in computer-based human assessment. Dordrecht, Kluwer Academic Publihers.

De Mauro, G. E. (1990). Effects of representation of gender group in the examinee population on the Mantel-Haenszel procedure. Texte présenté à la réunion annuelle de l'American Educational Research Association, Boston, MA. ED 318 747.

Donoghue, J. R., Holland, P. W., et Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. Dans P. W. Holland et H. Wainer (Eds.), Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Doolittle, A. E. (1983). The reliability of measuring differential item performance. Texte présenté à la réunion annuelle de l'American Educational Research Association, Montréal. ED 234 061.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. Applied Measurement in Education, 2(3), 217-233.

Dorans, N. J., et Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. Dans P. W. Holland et H. Wainer (Eds.), Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Dorans, N. J., et Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December, 1977: An application of the standardization approach. (Research Report No. 83-9). Princeton, NJ: Educational Testing Service.

Dorans, N. J., et Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23(4), 355-368.

Dorans, N. J., Potenza, M. T., et Ramsay, J. O. (1993). Smoothed standardization: A small sample DIF procedure. Texte inédit.

Dorans, N. J., Schmitt, A. P., et Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. Journal of Educational Measurement, 29(4), 309-319.

Dorans, N. J., Schmitt, A. P., et Bleistein, C. A. (1988). The standardization approach to assessing differential speededness. (Research Report, No. 88-31). Princeton, NJ: Educational Testing Service.

Engelhard, G., Anderson, D., et Gabrielson, S. (1990). An empirical comparison of Mantel-Haenszel and Rasch procedures for studying differential item functioning on teacher certification tests. Journal of Research and Development in Education, 23(3), 172-179.

Green, B. F. (1991). Differential item functioning: Techniques,findings and prospects. Texte présenté à la conférence intitulée «Modern theories for measurement: Issues and practices», Montebello.

Green, D. R., et Draper, J. F. (1972). Exploratory studies of bias in achievement tests. Texte présenté à la réunion annuelle de l'American Psychological Association, Honolulu, HA. ED 070 794.

Gutierrez, J. (1989). Characteristics of the distribution of the Mantel-Haenszel delta under different conditions of the null hypothesis: A Monte Carlo study. (Thèse de doctorat inédite, University of Ottawa).

Hambleton, R. K., et Jones, R. W. (1992). Comparison of empirical and judgemental methods for differential item functioning. Texte présenté à la réunion annuelle du National Council on Measurement in Education, San Francisco, CA.

Hambleton, R. K., et Rogers, H. J. (1988). Detecting biased test items: Comparison of the IRT area and Mantel-Haenszel methods. Texte présenté à la réunion annuelle de l'American Educational Research Association, New Orleans, LA.

Hambleton, R. K., et Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2(4), 313-334.

Hambleton, R. K., et Rogers, H. J. (1991). Evaluation of the plot method for identifying potentially biased test items. Dans P. L. Dann, S. H. Irvine et J. M. Collis (Eds.), Advances in computer-based human assessment. Dordrecht, Kluwer Academic Publishers.

Hambleton, R. K., Rogers, H. J., et Arrasmith, D. (1988). Identifying potentially biased test items: A comparison of the Mantel-Haenszel statistic and several item response theory methods. Texte présenté à la réunion annuelle de l'American Educational Research Association, San Francisco, CA.

Hambleton, R. K., et Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Nijhoff Publishing.

Hambleton, R. K., et Zaal, J. N. (Eds.). (1991). Advances in educational and psychological testing : Theory and applications. Boston, MA: Kluwer Academic Publishers.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. Applied Psychological Measurement, 9(2), 139 - 164.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8(4), 5-11.

Holland, P. W., et Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. Dans H. Wainer et H. I. Braun (Eds.), Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., et Thayer, D. T. (1986). Differential item performance and the Mantel-Haenszel procedure. Texte présenté à la réunion annuelle de l'American Educational Research Association, San Francisco, CA. ED 272 577.

Holland, P. W., et Wainer, H. (Eds.) (1993). Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Hoover, H. D., et Kolen, M. J. (1984). The reliability of six item bias indices. Applied Psychological Measurement, 8(2), 173-181.

Hulin, C. L., Drasgow, F., et Parsons, C. R. (1983). Item response theory. Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

Hunter, J. E. (1975). A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. Texte présenté à la National Institute of Education Conference on Test Bias, Annapolis, MD.

Ibrahim, A. K. (1992). Distribution and power of selected item bias indices: A Monte Carlo study. (Thèse de doctorat inédite. University of Ottawa).

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. Dans R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Ironson, G. H., et Subkovick, M.J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 16(4).

Kok, F. G., Mellenbergh, G. J., et Van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. Journal of Educational Measurement, 22(4), 295-303.

Laforge, H. (1891). Analyse multivariée pour les sciences sociales et biologiques avec applications des logiciels BMD, BMDP, SPSS, SAS. St-Laurent, Éditions Études Vivantes.

Linacre, J. M., et Wright, B. D. (1988). Item bias: Mantel-Haenszel and the Rasch model. Finnish Association of Mathematics and Science Education Research. Memorandum 39.

Linn, R. L., et Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction. Educational Measurement: Issues and Practice, 6(2), 13-17.

Liu, C. F., Dorans, N. J., et Ramsay, J. O. (1995). Smoothed standardization assessment of testlet level DIF on a Math-free response item type. Research Report RR-95-38. Educational Testing Service, Princeton, NJ.

Longford, N. T., Holland, P. W., et Thayer, D. E. (1993). Stability of the MH D DIF statistics across populations. Dans P. W. Holland et H. Wainer (Eds.), Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. Dans Y. H. Poortinga (Ed.), Basic problems in cross-cultural psychology. Amsterdam, Swets et Zeitlinger.

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

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

Loyd, B. (1984). Evaluation of log-linear models for detection of item bias: A comparison across samples. Texte présenté à la réunion du National Council on Measurement in Education, New Orleans, LA. ED 247 299.

Loyd, B. (1983). The effect of number of ability intervals on the stability of item bias detection. Texte présenté à la réunion de l'Eastern Educational Research Association. Baltimore, MD. ED 247 297.

Marascuilo, L. A., et Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. Journal of Educational Measurement, 18(4), 229-248.

Mazor, K. M., Clauser, B. E., et Hambleton R. K. (1991 a). Identification of non-uniform differential item functioning using a variation of the Mantel-Haenszel procedure. (Laboratory of Psychometric and Evaluative Research Report No. 227). Amherst, MA: University of Massachusetts.

Mazor, K. M., Clauser, B. E., et Hambleton R. K. (1991 b). The effect of sample size on the Mantel-Haenszel statistic. Texte présenté à la réunion annuelle du National Council on Measurement in Education, Chicago, IL. ED 331 877.

McCauley, C. D., et Mendoza, J. (1985). A simulation study of item bias using a two parameter item response model. Applied Psychological Measurement, 9(4), 389-400.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7(2), 105-118.

Mellenbergh, G. J. (1989). Item bias and item response theory. International Journal of Educational Research, 13, 127-143.

Mellenbergh, G. J., et Kok, F. G. (1991). Finding the biasing trait(s). Dans P. L. Dann, S. H. Irvine et J. M. Collis (Eds.), Advances in computer-based human assesment. Dordrecht, Kluwer Academic Publishers.

Merz, W. R., et Grossen, N. E. (1979). An empirical investigation of six methods for examining test item bias. (Technical Report No. Grant N1E-6-78-0067). National Institute of Education, California State University, Sacramento. ED 178-566

Miller, M. D., et Oshima, T. C. (1992). Effect of sample size, number of biased items and magnitude of bias on a two-stage item bias estimation method. Applied Psychological Measurement, 16(4), 381-388.

Moroney, M. J. (1970). Comprendre la statistique - vérités et mensonges des chiffres. Verviers, Marabout université.

Norušis, M. J. (1990). SPSS-X advanced statistics user's guide. Chicago, Il: SPSS Inc.

Ochieng, C. M. O. (1992). Examination of the effects of sample size, item discrimination, difficulty parameter and ability distribution on the logistic regression statistics and the Mantel-Haenszel statistic under diffferent conditions of the null hypothesis: A Monte Carlo study. (Thèse de maîtrise inédite. University of Ottawa).

Osterlind, S. J. (1983). Test item bias. Beverly Hills, CA: Sage Publications.

Pang, X. L., et Boss, M. W. (1993). The effects of sample size, item difficulty, and item discrimination on logistic regression item bias indices. Texte présenté à la réunion annuelle de l'American Educational Research Association, Atlanta, GA.

Pang, X. L., Tian, F., et Boss, M. W. (1994). Performance of MH and LR procedures over replications using real data. Texte présenté à la réunion annuelle de l'American Educational Research Association.

Park, D.-G., et Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. Applied Psychological Measurement, 14(2), 163-173.

Perlman, C. L., Bezruezko, N., Junker, L. K., Reynolds, A. J., Rice, W. K., et Schulz, E. M. (1988). Investigating the stability of four methods for estimating item bias. Texte présenté à la réunion annuelle du National Council on Measurement in Education, New Orleans, LA.

Plake, B. S., et Hoover, H. D. (1979-1980). An analytical method of identifying biased test items. The Journal of Experimental Education, 48(2), 153-154.

Poortinga, N. H. (Ed.). (1977). Basic problems in cross-cultural psychology. Amsterdam, Switz et Vitlinger.

Raju, N. S., Bode, R. K., et Larsen, V. S. (1989). An empirical assessment of Mantel-Haenszel statistic for studying differential item performance. Applied Measurement in Education, 2(1), 1-13.

Raju, N. S., Bode, R. K., et Larsen, V. S. (1988). An empirical comparison of the Mantel-Haenszel technique with the delta and modified-delta methods for studying differential item performance. Texte présenté à la réunion annuelle de l'American Educational Research Association, New Orleans, LA.

Raju, N. S., Drasgow, F., et Slinde, J. A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. Educational and Psychological Measurement, 53(2), 301-314.

Ramsay, J. O. (1989). A comparison of three simple test theory models. Psychometrika, 54(3), 487-599.

Ramsay, J. O. (1991 a). Kernel smoothing approaches to nonparametric item characteristic curve estimation. Psychometrika, 56(4), 611-630.

Ramsay, J. O. (1991 b). Maximum marginal likelihood estimation for semiparametric item anaysis. Psychometrika, 58(3), 365-379.

Ramsay, J. O. (1992 a). Some notes on the statistical analysis of tests. Montréal, McGill University.

Ramsay, J. O. (1992 b). TESTGRAF - A program for the graphical analysis of multiple choice test and questionnaire data. Montréal, McGill University.

Ramsay, J. O. (1993). TESTGRAF - A program for the graphical analysis of multiple choice test and questionnaire data. Montréal, McGill University.

Ramsay, J. O. (1995). TestGraF - A program for the graphical analysis of multiple choice test and questionnaire data. Montréal, McGill University.

Rivera, C., et Schmitt, A. P. (1988). A comparison of Hispanic and White students'omit patterns on the Scholastic Aptitude Test. (ETS Research Report No. 88-44). Princeton, NJ: Educational Testing Service.

Rogers, H. J., et Hambleton, R. K. (1994). MH : A FORTRAN V program to compute the Mantel-Haenszel statistic for detecting differential item functioning. Texte inédit. Amherst, MA: University of Massachusetts.

Rogers, H. J., et Swaminathan, H. (1990). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Texte présenté à la réunion annuelle de l'AERA, Boston, MA.

Roznowski, M., Tucker, L. R., et Humphreys, L. G. (1991). Three approaches to determining the dimensionality of binary items. Applied Psychological Measurement, 15(2), 109 - 127.

Rudner, L. M., Getson, P. R., et Knight, D. L. (1980 b). Biased item detection techniques. Journal of Educational Statistics, 5(3), 213-233.

Rudner, L. M., Getson, P. R., et Knight, D. L. (1980a). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17(1), 1-10.

Ryan, K. E. (1990). The performance of the Mantel-Haenszel procedure. Texte présenté à la réunion annuelle de l'American Educational Research Association, Boston, MA.

Ryan, K. E. (1991). The performance of the Mantel-Haenszel procedure across samples and matching criteria. Journal of Educational Measurement, 28(4), 325-337.

Scheuneman, J. D. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16(3), 143-152.

Scheuneman, J. D., et Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2(3), 255-275.

Schmitt, A. P. (1987). Unexpected differential item performance of Hispanics examinees. Dans A. P. Schmitt et N. J. Dorans (Eds.), Differential item functioning on the Scholastic Aptitude Test (Research Memorandum N0. 87-1). Princeton, NJ: Educational Testing Service.

Schmitt, A. P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. Journal of Educational Measurement, 25(1), 1-13.

Schmitt, A. P., et Bleistein, C. A. (1987). Factor affecting differential item functioning for Black examinees on Scholastic Aptitude Test analogy items. (ETS Research Report No, 87-23). Princeton, NJ: Educational Testing Service.

Schmitt, A. P., et Dorans, N. J. (Eds.). (1987). Differential item functioning on the Scholastic Aptitude Test (Research Memorandum N0. 87-1). Princeton, NJ: Educational Testing Service.

Schmitt, A. P., et Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. Journal of Educational Measurement, 27(1), 67-81.

Schmitt, A. P., Dorans, N. J., Crone, C. R., et Maneckshava, B. T. (1991). Differential speededness and item omit on the SAT. (ETS Research Report No. 91-50). Princeton, NJ: Educational Testing Service.

Schulz, E. M., Perlman, C., Rice, W. K., et Wright, B. D. (1989). An empirical comparison of Rash and Mantel-Haenszel procedures for assessing item bias. Texte présenté à la réunion annuelle du National Council of Measurement in Education, San Francisco, CA.

Seong, T. J., et Subkoviak, M. J. (1987). A comparative study of recently proposed item bias detection methods. Texte présenté à la réunion annuelle du National Council of Measurement in Education, Washington, DC. ED 281 883.

Shepard, L. A. (1982). Definitions of bias. Dans R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Shepard, L. A., Camilli, G., et Averill, M. (1981). Comparison of procedures for detecting test-item bias with internal and external criteria. Journal of Educational Statistics, 6(4), 317-375.

Shepard, L. A., Camilli, G., et Williams, D. M. (1984). Accounting for statistical artifacts in item bias reseach. Journal of Educational Statistics, 9(2), 93-128.

Shepard, L. A., Camilli, G., et Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22(2), 77-105.

Shermis, M. D. et St. George, R. (1990). Item bias in mathematics achievement: The progressive achievement tests for mathematics. Texte présenté à la réunion annuelle du National Council for Educational Measurement, Boston, Ma.

Sinnot, L. T. (1980). Differences in item performance across groups. (ETS Research Report 89-19). Princeton, NJ: Educational Testing Service.

Skaggs, G. et Lissitz, R. W. (1988). Consistency of selected item bias indices: Implications of another failure. Texte présenté à la réunion annuelle de l'AERA, New Orleans, LA.

Spray, J. A. (1989). Performance of three conditional DIF statistics in detecting differential item fuctioning on simulated test. (ACT Report series No. 89-7). IO: American College Testing Program.

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

Swaminathan, H., et Rogers, H. J. (1990 a). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Texte présenté à la réunion annuelle de l'American Educational Research Association, Boston, MA.

Swaminathan, H., et Rogers, H. J. (1990 b). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 24(4), 361-370.

Sykes, R. C., et Fitzpatrick A. R. (1990). Establising a Mantel-Haenszel alpha cutscore through a multiple method procedure. Texte présenté à la réunion annuelle de l'American Educational Research Association, Boston, MA.

Tian, F., Pang, X. L., et Boss, M. W. (1994). The consistency of the Mantel-Haenszel and logistic regression DIF identification procedures across sample size and over replications. Texte présenté à la réunion annuelle de l'American Educational Research Association.

Tian, F., Pang, X. L., et Boss, M. W. (1994). The effects of sample size and criterion variableon the identification of DIF by the Mantel-Haenszel and logistic regression procedures. Texte présenté à la réunion annuelle du National Council on Measurement in Education.

Van de Vijver, F. J. R., et Poortinga, Y. H. (1991). Testing across cultures. Dans R. K. Hambleton et J. N. Zaal (Eds.), Advances in educational and psychological testing: Theory and Applications. Boston, MA: Kluwer Academic Publishers.

Van der Flier, H., Mellenbergh, G. J., Ader, H. J., et Wijn, M. (1984). An iterative item bias detection method. Journal of Educational Measurement, 21(2), 131-145.

Wainer, H., et Braun H. I. (Eds.). (1988). Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Wright, D. J. (1986). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. Texte présenté à la réunion annuelle du National Council of Measurement in Education, San Francisco, CA.

Wright, D. J. (1987). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. Dans A. P. Schmitt et N. J. Dorans (Eds.), Differential item functioning on the Scholastic Aptitude Test (Research Memorandum No. 87-1) Princeton, NJ: Educational Testing Service.

Zimmerman, D. W., Williams, R. H., et Zumbo, B. D. (1993). Reliability of measurement and power of significance tests based on differences. Applied Psychological Measurement, 17(1), 1-9.

Zwick, R., et Ercikan, K. (1989). Analysis of differential item functioning in the NAEP History Assessment. Journal of Educational Measurement, 26(1), 55-66.

ANNEXE A

FIGURES ET TABLEAUX AYANT TRAIT À L'ÉTUDE DU TEST ET ILLUSTRATION DES DEUX TYPES DE FONCTIONNEMENT DIFFÉRENTIEL

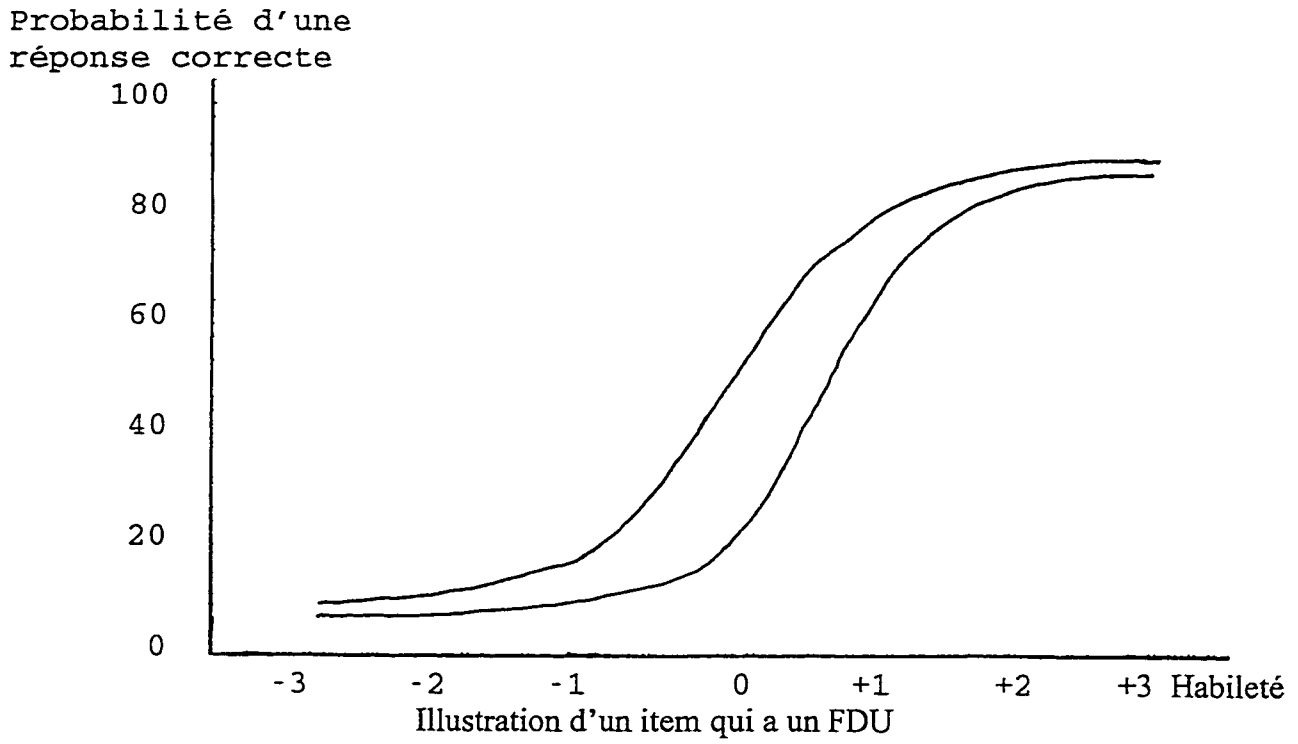


Illustration d'un item qui a un FDU

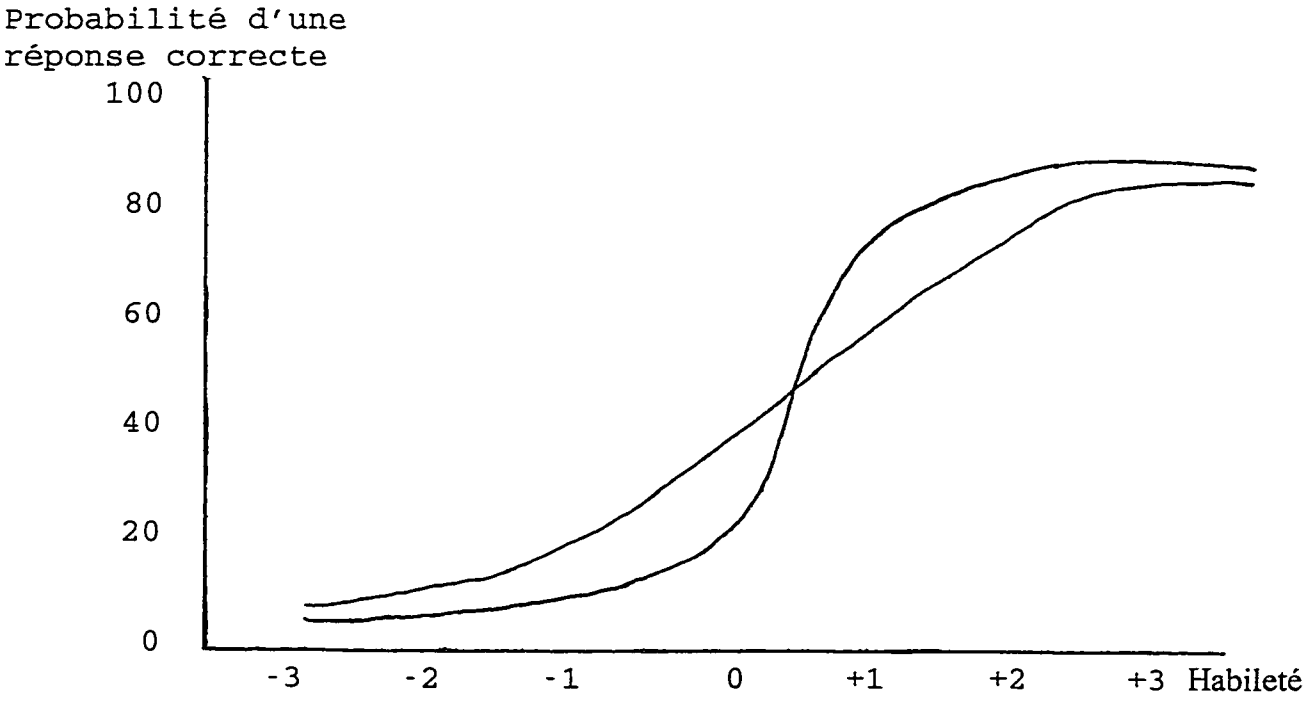


Illustration d'un item qui a un FDNU

Figure 1. Illustrations des courbes de réponses correctes en fonction du type de fonctionnement différentiel.

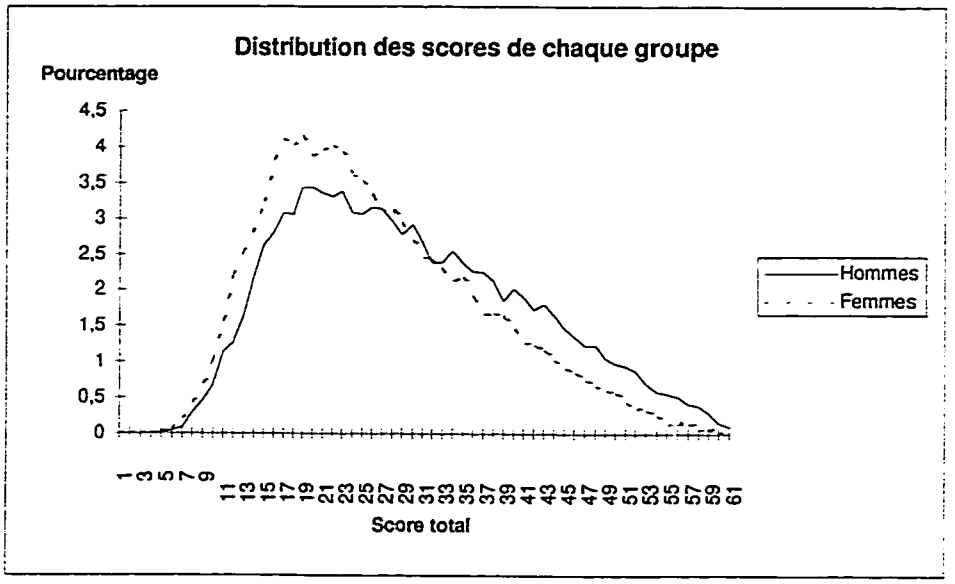


Figure 2. Distribution des scores au test de mathématiques pour les hommes (n = 16 521) et les femmes (n = 23 479).

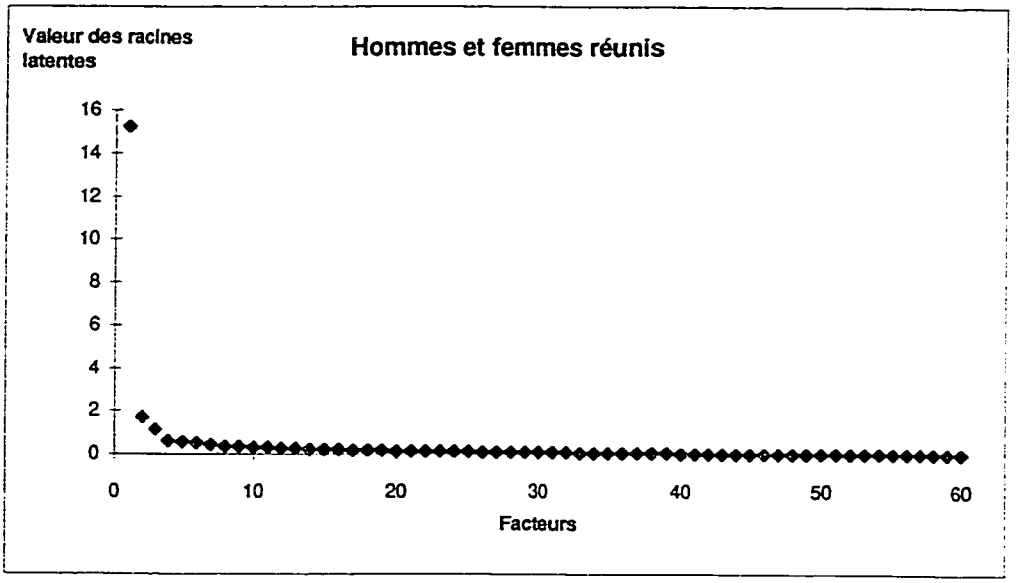
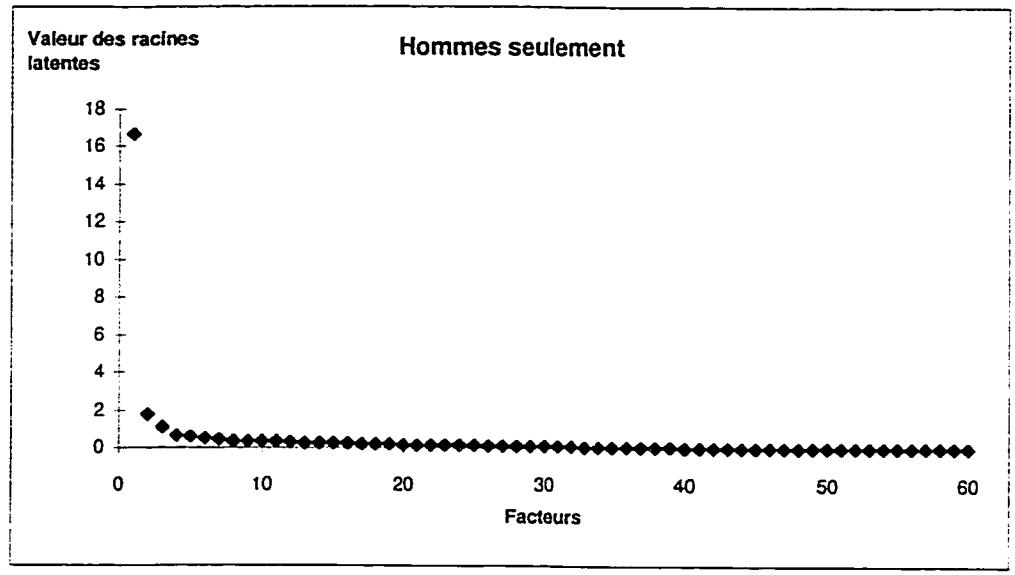


Figure 3a. Valeur des racines latentes en fonction de l'ordre des facteurs pour les racines latentes issues de matrices de corrélation tétrachorique (n = 40 000 sujets : 16 521 hommes et 23 479 femmes).



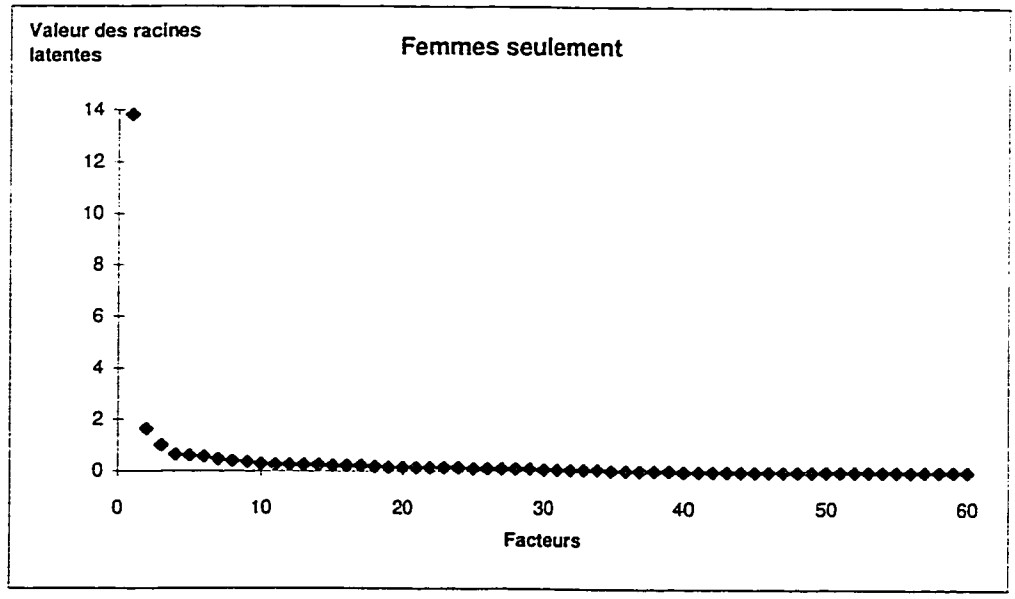


Figure 3b. Valeur des racines latentes en fonction de l'ordre des facteurs pour les racines latentes issues de matrices de corrélation tétrachorique (n = 40 000 sujets : 16 521 hommes et 23 479 femmes).

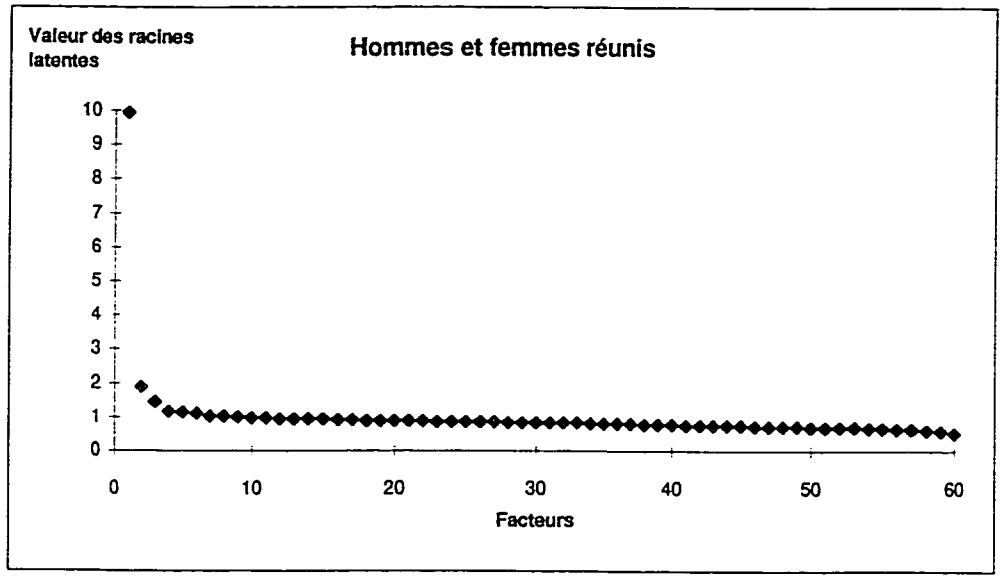
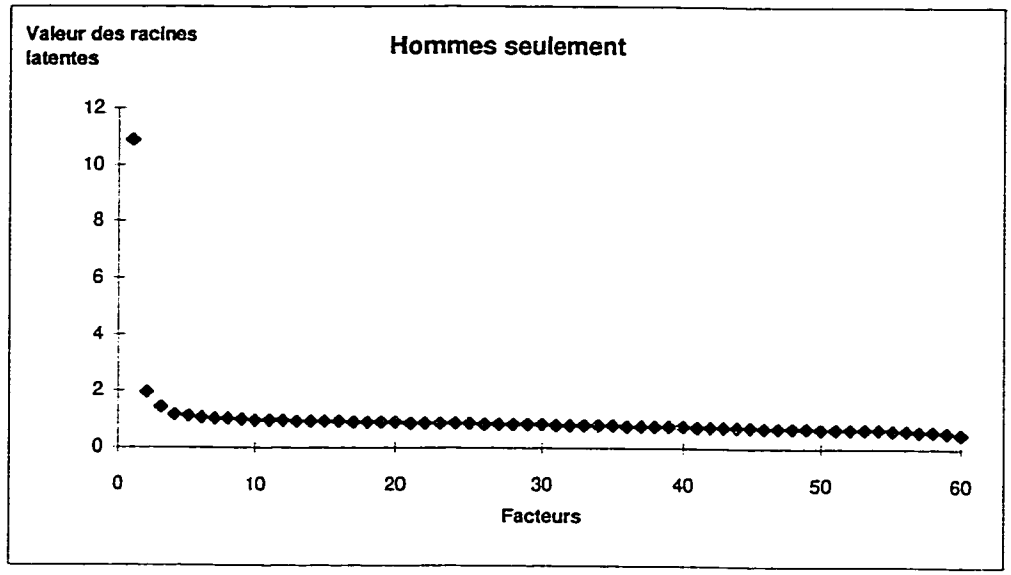


Figure 4a. Valeur des racines latentes en fonction de l'ordre des facteurs pour les racines latentes issues de matrices de corrélation phi (n = 40 000 sujets : 16 521 hommes et 23 479 femmes).



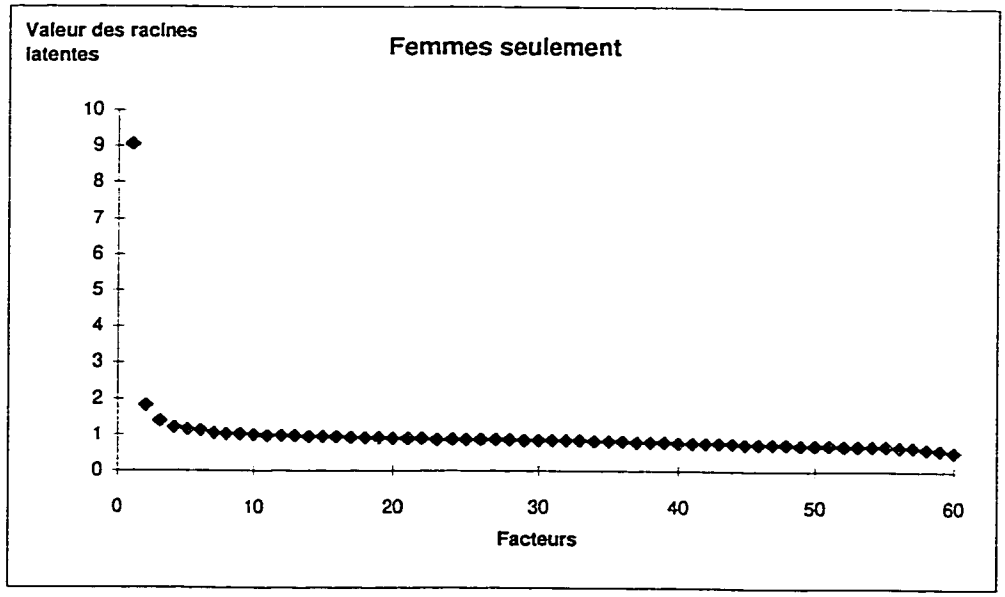


Figure 4b. Valeur des racines latentes en fonction de l'ordre des facteurs pour les racines latentes issues de matrices de corrélation phi (n = 40 000 sujets : 16 521 hommes et 23 479 femmes).

Tableau 1. Statistiques relatives aux racines latentes les plus grandes, rapport L et pourcentage de variance expliquée à partir d'analyses en composantes principales effectuées avec des coefficients de corrélation tétrachorique (N = 40 000 sujets : 16 521 hommes et 23 479 femmes).

Groupe	Racine latente			Rapport L		Pourcentage de variance expliquée
	Facteur*	Valeur	Différence	Valeur	Grandeur**	
Hommes	1	16,679	14,885	20,844	13,376	27,798
	2	1,794	0,714	1,558	0,105	2,991
	3	1,080	0,458	14,778	n/a	1,800
	4	0,622	0,310	n/a	n/a	1,037
Femmes	1	13,807	12,162	19,056	11,406	23,012
	2	1,646	0,638	1,671	0,090	2,742
	3	1,007	0,382	18,635	n/a	1,679
	4	0,625	0,021	n/a	n/a	1,042
Hommes et femmes	1	15,211	13,490	22,953	20,726	25,351
	2	1,720	0,588	1,107	0,039	2,867
	3	1,133	0,531	28,168	n/a	1,888
	4	0,602	0,019	n/a	n/a	1,003

* Seuls les facteurs dont la différence par rapport aux facteurs subséquents est supérieure à 0,100 sont indiqués.

** La grandeur du rapport L indique combien plus grand est le rapport par rapport au rapport subséquent.

Tableau 2. Répartition des items en fonction du domaine de contenu et de leur pondération sous le premier facteur (N = 40 000 sujets : 16 521 hommes et 23 479 femmes).

Pondération	Groupe	Domaine 1	Domaine 2	Domaine 3	Total
0,70 - 0,79	Hommes	3	0	0	3
	Femmes	3	0	0	3
	Hommes et femmes	3	0	0	3
0,60 - 0,69	Hommes	6	6	2	14
	Femmes	5	2	1	8
	Hommes et femmes	5	3	1	9
0,50 - 0,59	Hommes	6	4	6	16
	Femmes	5	5	4	14
	Hommes et femmes	6	5	6	17
0,40 - 0,49	Hommes	4	5	6	15
	Femmes	3	5	6	14
	Hommes et femmes	5	6	7	18
0,30 - 0,39	Hommes	4	2	3	9
	Femmes	6	3	4	13
	Hommes et femmes	3	3	2	8
0,00 - 0,29	Hommes	1	1	1	3
	Femmes	2	3	3	8
	Hommes et femmes	2	1	2	5

Nota bene : Domaine 1 : Algèbre fondamental.
Domaine 2 : Algèbre ou géométrie analytique.
Domaine 3 : Géométrie plane et trigonométrie.

Tableau 3. Répartition des items en fonction du domaine de contenu et de leur pondération sous le deuxième facteur (N = 40 000 sujets : 16 521 hommes et 23 479 femmes).

Pondération	Groupe	Domaine 1	Domaine 2	Domaine 3	Total
0,30 - 0,39	Hommes	2	0	1	3
	Femmes	3	0	1	4
	Hommes et femmes	3	0	1	4
0,20 - 0,29	Hommes	4	3	3	10
	Femmes	2	4	2	8
	Hommes et femmes	2	4	2	8
0,10 - 0,19	Hommes	2	1	3	6
	Femmes	2	0	2	4
	Hommes et femmes	2	0	3	5
0,00 - 0,09	Hommes	4	2	2	8
	Femmes	5	5	3	13
	Hommes et femmes	5	3	3	11
Valeurs négatives	Hommes	12	12	9	33
	Femmes	12	9	10	31
	Hommes et femmes	12	11	9	32

Nota bene : Domaine 1 : Algèbre fondamental.
Domaine 2 : Algèbre ou géométrie analytique.
Domaine 3 : Géométrie plane et trigonométrie.

Tableau 4. Relation entre le deuxième facteur du test de mathématiques et diverses variables (N = 60 items).

Pour les hommes seulement (n = 16 521 hommes)			
	Premier facteur	Deuxième facteur	Indices de difficulté
Deuxième facteur	-0,249		
Indices de difficulté	0,152	-0,925	
Taux d'abstention	-0,294	0,790	-0,788
Pour les femmes seulement (n = 23 479 femmes)			
Deuxième facteur	-0,375		
Indices de difficulté	0,307	-0,907	
Taux d'abstention	-0,427	0,815	-0,782
Pour les hommes et les femmes réunis (n = 40 000 personnes)			
Deuxième facteur	-0,352		
Indices de difficulté	0,230	-0.918	
Taux d'abstention	-0,358	0,823	-0,789

Nota bene : La corrélation entre les pondérations de chaque groupe pris séparément est de 0,946 pour le premier facteur et de 0,974 pour le deuxième facteur.

Tableau 5. Résultats de l'application du test T de Stout (1987) pour vérifier si le test de mathématiques de l'ACT peut être considéré comme unidimensionnel (N = 40 000 sujets : 16 521 hommes et 23 479 femmes).

Hypothèse vérifiée	**Groupe**	**T**	**p**	**T'**	**p**
Deuxième facteur distinct du premier * (Items qui se différencient le plus)	Hommes	-0,500	0,692	-0,452	0,674
	Femmes	0,179	0,429	0,390	0,348
	Hommes et femmes	-1,666	0,952	-1,894	0,971
Présence d'un facteur temps ** (Derniers items)	Hommes	2,063	0,020	2,251	0,012
	Femmes	1,298	0,097	2,038	0,020
	Hommes et femmes	0,336	0,611	0,102	0,541

* Items qui ont des pondérations élevées sous le deuxième facteur et des pondérations faibles sous le premier.

** Items qui se suivent et dont le taux d'abstention est supérieur à 0,05.

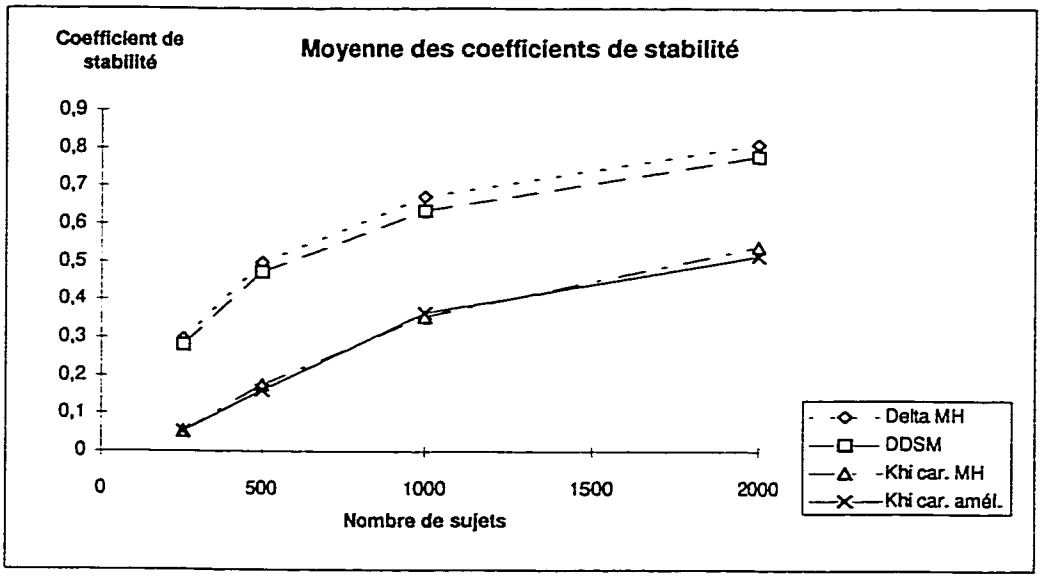
*** Légende : T : Résultat du test T de Stout (1987)
T' : Résultat du test T de Stout modifé par Nandakumar (1992)
P : Probabilité de rejet de l'hypothèse nulle, à savoir, que le test est multidimensionnel.

Tableau 6. Probabilité de sélection des sujets pour chaque grandeur d'échantillon et pourcentage de sujets communs à deux échantillons de même taille (N = 40 000 sujets, 16 521 hommes et 23 479 femmes).

Grandeur de l'échantillon	**Composition de l'échantillon**	**Probabilité de sélection des sujets**	**Pourcentage de sujets communs à deux échantillons (moyenne)**
2000 sujets	1000 hommes	0,061	0,061
	1000 femmes	0,043	0,043
	2000 personnes	0,052	0,052
1000 sujets	500 hommes	0,030	0,030
	500 femmes	0,021	0,021
	1000 personnes	0,026	0,026
500 sujets	250 hommes	0,015	0,015
	250 femmes	0,011	0,011
	500 personnes	0,013	0,013
250 sujets	125 hommes	0,008	0,008
	125 femmes	0,005	0,005
	250 personnes	0,006	0,006

ANNEXE B

FIGURES ET TABLEAUX AYANT TRAIT AUX ANALYSES DE FONCTIONNEMENT DIFFÉRENTIEL EFFECTUÉES



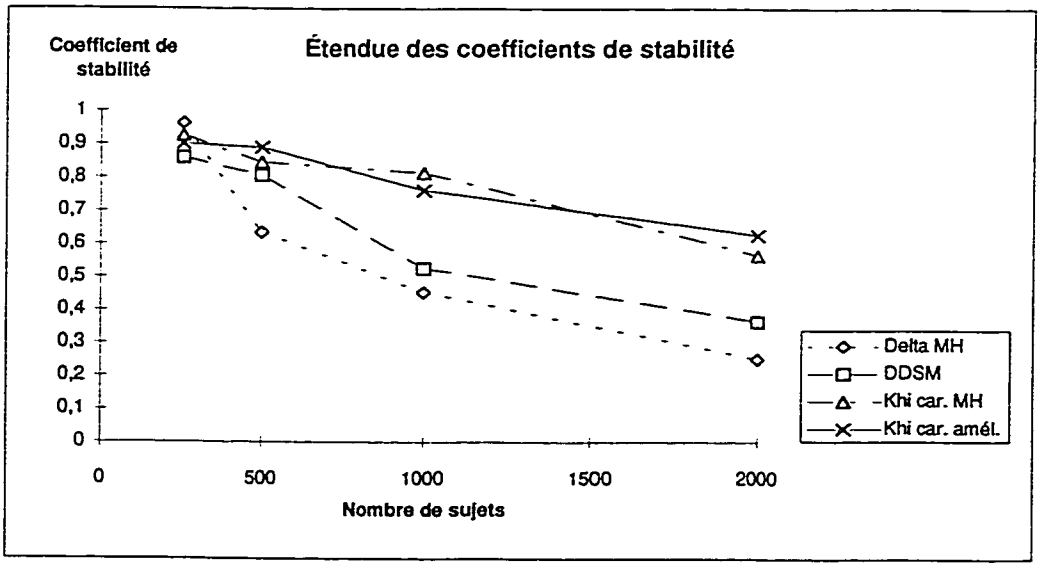


Figure 5. Relation entre la moyenne des coefficients de stabilité ou l'étendue de leurs distributions et la grandeur des échantillons (coefficients de corrélation de rangs ou coefficients de corrélation de Pearson, selon les indices).

Tableau 7a. Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour le delta de Mantel-Haenszel (N = 100 échantillons).

Item	250 sujets						
	Minimum	Maximum	Médiane	1er quartile	3e quartile	Moyenne	Écart-type
Item1	-2,44	3,03	0,14	-0,465	0,735	0,1602	1,014216
Item2	-3,2	1,39	-0,49	-1,0325	-0,01	-0,5491	0,851764
Item3	-2,53	2,56	0,23	-0,4025	0,74	0,1569	0,924616
Item4	-1,69	3,11	0,795	0,2525	1,3075	0,7706	0,815679
Item5	-1,37	2,28	0,415	-0,0375	0,92	0,4313	0,73124
Item6	-2,2	1,7	-0,605	-1,24	0,075	-0,5815	0,845069
Item7	-3,38	0,68	-0,87	-1,4475	-0,485	-0,9814	0,835373
Item8	-2,28	1,48	-0,275	-0,85	0,1525	-0,3319	0,746076
Item9	-0,81	2,45	0,8	0,4125	1,2325	0,8109	0,696306
Item10	-1,13	1,91	0,405	-0,105	0,7825	0,3327	0,674604
Item11	-1,99	1,64	-0,205	-0,6425	0,24	-0,2233	0,708142
Item12	-1,55	2,01	0,275	-0,2025	0,7625	0,2867	0,747964
Item13	-2,57	1,33	-0,41	-0,8675	0,11	-0,4091	0,7727
Item14	-1,97	3,08	1,12	0,6825	1,79	1,1716	0,869515
Item15	-2,67	1,45	-0,47	-1,21	0,025	-0,5576	0,883169
Item16	-1,62	2,82	0,69	0,135	1,085	0,5866	0,7689
Item17	-2,36	0,78	-1,02	-1,59	-0,5725	-1,025	0,73419
Item18	-0,83	2,43	0,58	0,2225	0,99	0,5954	0,631504
Item19	-0,55	3,93	1,405	0,8675	1,9	1,3844	0,799935
Item20	-2,09	1,82	-0,095	-0,4325	0,33	-0,0726	0,707443
Item21	-1,72	2,43	0,43	-0,0925	0,86	0,3492	0,722459
Item22	-1,89	2,28	0,485	-0,0325	1,02	0,4778	0,757157
Item23	-2,08	1,22	-0,67	-1,2725	-0,2975	-0,7334	0,710338
Item24	-2,18	2,66	0,065	-0,445	0,43	0,0374	0,810535
Item25	-1,44	2,43	0,225	-0,24	0,7675	0,2931	0,738316
Item26	-1,74	2,34	0,125	-0,2625	0,5425	0,1462	0,705689
Item27	-1,73	1,84	-0,015	-0,475	0,415	-0,0551	0,672707
Item28	-2,77	1,8	-0,395	-0,9225	0,1475	-0,4441	0,860277
Item29	-1,86	2,72	0,88	0,33	1,425	0,8752	0,85387
Item30	-1,85	1,59	0,16	-0,345	0,7025	0,1354	0,760795
Item31	-1,92	1,94	-0,125	-0,7625	0,28	-0,2312	0,779152
Item32	-2,31	0,6	-0,66	-1,06	-0,1875	-0,6799	0,664841
Item33	-1,72	2,21	0,35	-0,2925	0,7575	0,2382	0,793702
Item34	-2,32	1,65	-0,625	-1,135	-0,1475	-0,6058	0,777694
Item35	-1,75	1,77	-0,305	-0,7525	0,3	-0,2098	0,743496
Item36	-1,72	1,74	0,23	-0,175	0,8375	0,2365	0,760812
Item37	-1,23	2,4	0,72	0,16	1,26	0,7114	0,753971
Item38	-2,96	2,12	-0,79	-1,445	-0,2375	-0,8501	0,851858
Item39	-1,5	1,91	0,29	-0,15	0,77	0,2825	0,720587
Item40	-2,04	1,55	0,155	-0,4525	0,55	0,0231	0,764678
Item41	-1,79	3,1	0,485	-0,05	1,0075	0,4853	0,828607
Item42	-2,31	2,17	0,075	-0,4025	0,57	0,0545	0,835214
Item43	-2,8	1,48	-0,295	-0,8825	0,375	-0,3029	0,825598
Item44	-2,13	1,94	-0,07	-0,595	0,485	-0,1074	0,799189
Item45	-2,48	1,49	-0,31	-0,6875	0,22	-0,2778	0,754814
Item46	-2,37	4,16	0,345	-0,19	0,98	0,4086	0,948611
Item47	-1,95	1,81	0,115	-0,27	0,595	0,1284	0,698162
Item48	-2,69	0,9	-0,535	-1,09	-0,075	-0,6134	0,77613
Item49	-1,95	2,23	0,205	-0,565	0,7	0,0793	0,877326
Item50	-1,31	1,98	0,11	-0,4425	0,675	0,1313	0,711562
Item51	-1,32	2,72	0,46	-0,0325	1,065	0,4826	0,845411
Item52	-3,41	0,87	-0,945	-1,355	-0,415	-0,8692	0,828841
Item53	-2,46	2,58	-0,035	-0,695	0,4625	-0,0765	0,871736
Item54	-1,91	2,18	0,165	-0,58	0,835	0,1495	0,917122
Item55	-2,61	2,89	-0,27	-0,8325	0,1775	-0,2721	0,849536
Item56	-3,01	2,13	-0,375	-0,94	0,33	-0,3161	0,986055
Item57	-3,12	3,64	0,58	-0,1525	1,2675	0,5749	1,13607
Item58	-3,18	0,98	-1,185	-1,8425	-0,6	-1,2222	0,938753
Item59	-3,06	3,02	0,33	-0,3925	1,045	0,312	1,147685
Item60	-1,95	3,06	0,22	-0,54	0,905	0,2723	1,111006

Tableau 7a. (suite) Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour le delta de Mantel-Haenszel (N = 100 échantillons).

Item			500 sujets				
	Minimum	Maximum	Médiane	1er quartile	3e quartile	Moyenne	Écart-type
Item1	-0,93	2,17	0,265	-0,155	0,81	0,2917	0,665068
Item2	-2,48	0,92	-0,555	-1,03	-0,19	-0,5764	0,66319
Item3	-1,09	1,33	0,09	-0,285	0,4575	0,1002	0,529118
Item4	-1,2	2,07	0,77	0,4275	1,0875	0,7555	0,575936
Item5	-0,83	1,46	0,435	-0,005	0,75	0,3904	0,467423
Item6	-2,39	0,74	-0,69	-1,195	-0,2175	-0,6962	0,654184
Item7	-2,21	0,44	-0,76	-1,0825	-0,44	-0,7499	0,496292
Item8	-1,35	0,8	-0,23	-0,605	0,0225	-0,2801	0,459331
Item9	-0,28	1,86	0,785	0,485	1,25	0,8342	0,499685
Item10	-0,83	2,25	0,44	0,1625	0,8025	0,4922	0,546405
Item11	-1,3	0,71	-0,14	-0,5325	0,09	-0,2225	0,442325
Item12	-0,77	1,51	0,315	-0,06	0,79	0,337	0,552435
Item13	-1,49	0,99	-0,305	-0,6325	0	-0,3306	0,482606
Item14	-0,09	2,22	1,06	0,7525	1,57	1,1012	0,565198
Item15	-1,54	1,17	-0,38	-0,78	-0,1375	-0,4325	0,520535
Item16	-0,75	2	0,655	0,3275	0,93	0,6141	0,499539
Item17	-2,26	-0,01	-0,935	-1,3625	-0,66	-0,9851	0,520389
Item18	-0,64	1,79	0,46	0,12	0,71	0,454	0,464238
Item19	-0,24	2,8	1,475	1,075	1,7425	1,4438	0,554904
Item20	-1,1	0,9	0	-0,18	0,32	0,0358	0,424107
Item21	-0,55	1,88	0,35	0,0375	0,7725	0,3953	0,492996
Item22	-0,74	1,47	0,45	0,1425	0,8125	0,4408	0,509666
Item23	-1,93	0,35	-0,74	-1,09	-0,48	-0,7958	0,440482
Item24	-0,85	1,18	0,155	-0,205	0,4425	0,1219	0,459428
Item25	-0,9	1,27	0,185	-0,1725	0,59	0,1696	0,476044
Item26	-1,15	1,49	0,09	-0,195	0,4525	0,0922	0,545725
Item27	-1,32	1,56	-0,125	-0,4075	0,165	-0,1095	0,480936
Item28	-1,35	1,02	-0,315	-0,6825	0,045	-0,304	0,520206
Item29	-0,7	2,14	0,87	0,52	1,2525	0,8707	0,552342
Item30	-1,06	1,3	-0,02	-0,2475	0,3225	0,0143	0,446679
Item31	-1,31	1,19	-0,175	-0,4825	0,075	-0,1529	0,488372
Item32	-2,28	0,66	-0,675	-1,0625	-0,255	-0,6601	0,607245
Item33	-1,12	1,48	0,155	-0,18	0,57	0,1603	0,512344
Item34	-2,05	0,4	-0,605	-1,0025	-0,27	-0,6114	0,489068
Item35	-1,33	1,06	-0,18	-0,5025	0,11	-0,1665	0,469012
Item36	-1,14	1,51	0,28	-0,1125	0,57	0,2278	0,50119
Item37	-0,99	2,31	0,545	0,3275	0,8925	0,5721	0,504924
Item38	-2,49	0,03	-0,97	-1,2925	-0,6875	-0,975	0,50053
Item39	-1,03	1,38	0,29	-0,13	0,6275	0,2697	0,537971
Item40	-0,87	1,2	0,15	-0,13	0,39	0,1457	0,394757
Item41	-0,63	2,05	0,58	0,29	0,97	0,6132	0,553466
Item42	-0,88	1,04	0,145	-0,23	0,4125	0,0849	0,45364
Item43	-1,59	1,03	-0,215	-0,4825	0,12	-0,2186	0,525755
Item44	-1,17	1,06	-0,195	-0,5225	0,07	-0,1973	0,460577
Item45	-1,56	1,26	-0,19	-0,57	0,1625	-0,19	0,528594
Item46	-1,57	2,14	0,475	0,1225	0,8725	0,5005	0,59329
Item47	-1,24	1,43	0,035	-0,35	0,35	0,0202	0,505749
Item48	-1,77	0,3	-0,61	-0,8925	-0,25	-0,5906	0,484807
Item49	-1,27	1,38	-0,04	-0,32	0,4925	0,056	0,569719
Item50	-1,19	1,56	0,185	-0,165	0,48	0,1686	0,502979
Item51	-1,31	1,64	0,47	0,055	0,8525	0,4422	0,616203
Item52	-2,23	0,63	-0,825	-1,1925	-0,495	-0,8215	0,567805
Item53	-1,91	1,29	-0,24	-0,66	0,1625	-0,2492	0,584938
Item54	-1,43	1,61	-0,03	-0,38	0,2925	-0,0256	0,572465
Item55	-1,76	1,09	-0,455	-0,7825	-0,105	-0,459	0,56468
Item56	-2,1	1,32	-0,385	-0,8225	0,04	-0,3782	0,661948
Item57	-1,29	2,51	0,66	0,19	1,0625	0,629	0,711435
Item58	-2,33	1,21	-1,195	-1,5925	-0,935	-1,183	0,579382
Item59	-1,46	1,86	0,19	-0,2725	0,6575	0,2082	0,635402
Item60	-1,78	1,43	0,035	-0,4625	0,46	-0,0457	0,701252

Tableau 7a. (suite) Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour le delta de Mantel-Haenszel (N = 100 échantillons).

Item	1000 sujets						
	Minimum	Maximum	Médiane	1er quartile	3e quartile	Moyenne	Écart-type
Item1	-0,88	1,27	0,315	-0,0425	0,5	0,2373	0,45485
Item2	-1,44	0,45	-0,47	-0,7725	-0,24	-0,5057	0,411291
Item3	-0,72	1,01	0,035	-0,23	0,33	0,0512	0,392501
Item4	-0,09	1,64	0,78	0,535	1	0,7517	0,377129
Item5	-0,59	1,35	0,31	0,1	0,5675	0,3288	0,341456
Item6	-1,54	0,43	-0,725	-0,9325	-0,4375	-0,6754	0,386924
Item7	-1,63	0,53	-0,635	-0,9375	-0,4	-0,6749	0,377213
Item8	-1,13	0,61	-0,275	-0,48	0,01	-0,2374	0,353798
Item9	-0,13	1,55	0,685	0,4625	0,96	0,7109	0,360034
Item10	-0,43	1,68	0,485	0,2075	0,775	0,4818	0,38172
Item11	-0,96	0,48	-0,285	-0,5	0,02	-0,252	0,346745
Item12	-0,5	1,38	0,33	0,0825	0,5625	0,3149	0,352301
Item13	-1,08	0,82	-0,375	-0,58	-0,165	-0,3438	0,359452
Item14	0,17	2,28	1,095	0,82	1,3425	1,1104	0,414787
Item15	-1,11	0,67	-0,455	-0,7125	-0,1575	-0,4437	0,352709
Item16	-0,54	1,35	0,615	0,35	0,8875	0,6229	0,34451
Item17	-1,99	-0,14	-1,07	-1,31	-0,85	-1,0691	0,36948
Item18	-0,11	1,26	0,48	0,29	0,665	0,4962	0,293972
Item19	0,53	2,83	1,42	1,2175	1,725	1,435	0,391956
Item20	-0,65	0,83	-0,03	-0,2	0,2225	0,021	0,309945
Item21	-0,49	1,24	0,44	0,2475	0,69	0,4611	0,328498
Item22	-0,61	1,07	0,49	0,22	0,695	0,4406	0,363323
Item23	-1,68	0,13	-0,845	-0,98	-0,5625	-0,7815	0,328386
Item24	-0,67	0,84	0,045	-0,15	0,22	0,0396	0,309773
Item25	-0,59	0,99	0,16	-0,0925	0,3525	0,1285	0,329969
Item26	-0,87	0,76	0,015	-0,2325	0,22	0,0076	0,358282
Item27	-0,72	0,7	-0,135	-0,275	0,07	-0,102	0,280447
Item28	-1,05	0,53	-0,33	-0,515	-0,035	-0,2959	0,320996
Item29	-0,9	1,83	0,945	0,6775	1,19	0,9256	0,417889
Item30	-0,77	1,04	0,07	-0,15	0,3025	0,0768	0,353727
Item31	-1,13	0,71	-0,29	-0,49	-0,08	-0,2563	0,358168
Item32	-1,77	0,36	-0,625	-0,9175	-0,445	-0,6821	0,365608
Item33	-0,45	0,79	0,195	-0,08	0,3975	0,1768	0,305253
Item34	-1,62	0,2	-0,69	-0,945	-0,42	-0,6884	0,373157
Item35	-1,38	0,84	-0,08	-0,2825	0,12	-0,0966	0,346425
Item36	-0,61	1,21	0,175	-0,0425	0,3525	0,1504	0,33366
Item37	-0,3	1,69	0,62	0,3475	0,85	0,6069	0,358001
Item38	-2,06	-0,15	-0,875	-1,075	-0,635	-0,8981	0,358416
Item39	-0,29	1,01	0,325	0,0775	0,5425	0,3174	0,300466
Item40	-0,57	0,92	0,085	-0,09	0,32	0,1134	0,292803
Item41	-0,39	1,76	0,5	0,2675	0,8025	0,5425	0,401644
Item42	-0,57	0,73	0,08	-0,1325	0,28	0,0948	0,296857
Item43	-0,84	0,38	-0,225	-0,455	0,08	-0,208	0,329214
Item44	-1,18	0,75	-0,1	-0,405	0,2025	-0,106	0,386646
Item45	-1,22	0,56	-0,2	-0,45	0,01	-0,2202	0,373947
Item46	-0,59	1,24	0,49	0,15	0,68	0,4313	0,385032
Item47	-0,52	1,05	0,105	-0,12	0,3725	0,1239	0,335445
Item48	-1,67	0,39	-0,57	-0,785	-0,3225	-0,5801	0,385585
Item49	-0,94	1,03	-0,01	-0,195	0,28	0,005	0,385687
Item50	-0,85	1,05	0,245	-0,035	0,42	0,1787	0,320412
Item51	-0,59	1,46	0,35	0,1275	0,58	0,349	0,430066
Item52	-1,79	0,02	-0,795	-1,0725	-0,5775	-0,8408	0,387702
Item53	-1,16	0,87	-0,23	-0,525	0,0125	-0,2273	0,427532
Item54	-0,76	1,22	-0,005	-0,35	0,2475	-0,0022	0,420189
Item55	-1,24	0,7	-0,325	-0,6375	-0,035	-0,3254	0,398866
Item56	-1,43	1,02	-0,31	-0,6725	-0,01	-0,3428	0,468822
Item57	-0,3	1,99	0,59	0,285	0,9775	0,6285	0,474645
Item58	-1,96	-0,25	-1,225	-1,45	-1	-1,2068	0,358036
Item59	-0,67	1,3	0,28	0,0625	0,495	0,2711	0,37533
Item60	-0,8	1,08	-0,05	-0,3	0,27	0,0116	0,440458

Tableau 7a. (suite) Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour le delta de Mantel-Haenszel (N = 100 échantillons).

Item	2000 sujets						
	Minimum	Maximum	Médiane	1er quartile	3e quartile	Moyenne	Écart-type
Item1	-0,37	1,07	0,165	-0,0125	0,41	0,1973	0,312658
Item2	-1,39	0,14	-0,515	-0,71	-0,3375	-0,5231	0,313832
Item3	-0,72	0,76	0,09	-0,075	0,3025	0,1068	0,259225
Item4	0,01	1,37	0,785	0,57	0,95	0,7739	0,272389
Item5	-0,24	1,07	0,275	0,145	0,4925	0,3149	0,27173
Item6	-1,12	-0,03	-0,6	-0,77	-0,3975	-0,5889	0,269571
Item7	-1,3	-0,09	-0,715	-0,86	-0,575	-0,7025	0,232141
Item8	-0,87	0,34	-0,295	-0,44	-0,115	-0,2784	0,267694
Item9	0,08	1,33	0,69	0,5575	0,825	0,6824	0,221479
Item10	-0,08	1,15	0,45	0,3	0,6025	0,4674	0,234782
Item11	-0,79	0,42	-0,27	-0,3725	-0,08	-0,242	0,227241
Item12	-0,42	0,89	0,315	0,18	0,4425	0,3213	0,226874
Item13	-0,92	0,27	-0,335	-0,51	-0,2075	-0,3534	0,219523
Item14	0,57	1,82	1,12	0,975	1,29	1,1274	0,23777
Item15	-1,07	0,03	-0,43	-0,57	-0,26	-0,4264	0,233962
Item16	0	1,39	0,59	0,4175	0,77	0,6006	0,24541
Item17	-1,63	-0,09	-1,035	-1,26	-0,8375	-1,0387	0,269757
Item18	0,08	1,23	0,54	0,39	0,6825	0,5324	0,238713
Item19	0,84	2,18	1,425	1,28	1,5925	1,4325	0,240729
Item20	-0,57	0,58	-0,035	-0,1725	0,085	-0,0276	0,215365
Item21	-0,45	0,89	0,39	0,24	0,5125	0,3741	0,234478
Item22	-0,15	1,03	0,49	0,3375	0,64	0,483	0,252505
Item23	-1,33	-0,27	-0,805	-0,96	-0,6275	-0,7799	0,231008
Item24	-0,77	0,52	-0,03	-0,195	0,1425	-0,0255	0,238863
Item25	-0,29	0,77	0,155	0,0175	0,3025	0,1547	0,211524
Item26	-0,4	0,67	0,03	-0,11	0,17	0,0399	0,227825
Item27	-0,68	0,43	-0,08	-0,2125	0,07	-0,0703	0,209538
Item28	-0,73	0,4	-0,295	-0,4525	-0,14	-0,2925	0,226128
Item29	0,06	1,55	0,845	0,6575	1,0225	0,851	0,263391
Item30	-0,52	0,78	0,1	-0,0725	0,295	0,0969	0,276423
Item31	-0,75	0,44	-0,24	-0,385	-0,0775	-0,2279	0,223595
Item32	-1,35	-0,22	-0,68	-0,8625	-0,5175	-0,6859	0,234771
Item33	-0,37	0,62	0,18	0,0475	0,34	0,1769	0,203949
Item34	-1,29	-0,02	-0,665	-0,86	-0,5075	-0,6724	0,260175
Item35	-0,67	0,53	-0,08	-0,2225	0,0825	-0,0859	0,239777
Item36	-0,6	0,63	0,12	-0,06	0,2425	0,0993	0,215867
Item37	0,25	1,21	0,64	0,4875	0,8225	0,6527	0,223195
Item38	-1,58	-0,37	-0,875	-1,0925	-0,745	-0,9061	0,246191
Item39	-0,43	0,78	0,29	0,1075	0,44	0,2829	0,252316
Item40	-0,5	0,61	0,01	-0,11	0,1825	0,0394	0,224643
Item41	-0,14	1,25	0,565	0,355	0,7825	0,5635	0,276852
Item42	-0,38	0,62	0,09	-0,0825	0,2325	0,0947	0,216102
Item43	-0,87	0,6	-0,21	-0,3725	0,0175	-0,183	0,284429
Item44	-0,69	0,48	-0,14	-0,2625	0,04	-0,1222	0,241283
Item45	-0,71	0,27	-0,22	-0,415	-0,035	-0,2239	0,24202
Item46	-0,42	1,36	0,43	0,2275	0,58	0,4192	0,276492
Item47	-0,34	0,87	0,17	0,005	0,3	0,1612	0,221757
Item48	-1,29	-0,04	-0,575	-0,7475	-0,3775	-0,5625	0,263168
Item49	-0,57	0,63	0,115	-0,0725	0,26	0,1033	0,257321
Item50	-0,35	0,68	0,14	0,01	0,2925	0,141	0,208625
Item51	-0,27	1,11	0,395	0,22	0,56	0,4063	0,275533
Item52	-1,52	-0,27	-0,815	-0,97	-0,67	-0,8214	0,239368
Item53	-0,88	0,35	-0,315	-0,4575	-0,1375	-0,3002	0,260241
Item54	-0,73	0,99	0,045	-0,1625	0,295	0,0649	0,327453
Item55	-0,86	0,34	-0,325	-0,5225	-0,17	-0,3316	0,261556
Item56	-1,12	0,55	-0,325	-0,5825	-0,1475	-0,3613	0,348038
Item57	-0,1	1,53	0,735	0,51	0,9375	0,7257	0,33525
Item58	-1,9	-0,56	-1,2	-1,3725	-0,9975	-1,2063	0,291623
Item59	-0,34	1,12	0,225	-0,035	0,3825	0,1996	0,293505
Item60	-1	0,71	0,055	-0,1975	0,31	0,0333	0,353299

Tableau 7b. Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour la différence de difficulté standardisée modifiée (N = 100 échantillons).

Item	250 sujets						
	Minimum	Maximum	Médiane	1er quartile	3e quartile	Moyenne	Écart-type
Item1	-0,154	0,177	0,0155	-0,02425	0,0555	0,01559	0,065817
Item2	-0,278	0,102	-0,034	-0,075	0,01	-0,03681	0,068891
Item3	-0,145	0,168	0,0275	-0,02575	0,0675	0,02464	0,067134
Item4	-0,093	0,179	0,051	0,01175	0,09025	0,0495	0,059587
Item5	-0,101	0,167	0,052	0,00925	0,0895	0,04816	0,060215
Item6	-0,172	0,114	-0,027	-0,084	0,0035	-0,03641	0,062101
Item7	-0,215	0,138	-0,0795	-0,1155	-0,02825	-0,07053	0,070095
Item8	-0,178	0,136	-0,032	-0,0725	0,02825	-0,02274	0,072708
Item9	-0,06	0,2	0,0705	0,03125	0,10125	0,06797	0,057036
Item10	-0,107	0,202	0,0315	-0,01325	0,077	0,03356	0,065034
Item11	-0,165	0,135	-0,0245	-0,05575	0,02825	-0,01712	0,063868
Item12	-0,142	0,154	0,012	-0,02325	0,06025	0,01578	0,05725
Item13	-0,206	0,122	-0,026	-0,07775	0,01775	-0,03084	0,071461
Item14	-0,072	0,26	0,075	0,04275	0,10675	0,07584	0,052979
Item15	-0,265	0,084	-0,045	-0,09575	0,00375	-0,05115	0,07492
Item16	-0,122	0,162	0,046	-0,0015	0,08	0,04009	0,060847
Item17	-0,205	0,101	-0,079	-0,1255	-0,027	-0,07771	0,069174
Item18	-0,068	0,169	0,0505	0,01675	0,08125	0,0473	0,051886
Item19	-0,027	0,258	0,106	0,063	0,13425	0,10144	0,052124
Item20	-0,218	0,174	-0,0055	-0,03325	0,04	0,00043	0,067126
Item21	-0,151	0,197	0,031	-0,032	0,0575	0,01812	0,065039
Item22	-0,073	0,191	0,0355	-0,00025	0,07	0,0344	0,056949
Item23	-0,226	0,142	-0,0775	-0,1225	-0,0285	-0,07265	0,070655
Item24	-0,152	0,162	0,008	-0,03325	0,04275	0,00493	0,060425
Item25	-0,151	0,193	0,028	-0,01325	0,0605	0,02442	0,067432
Item26	-0,144	0,188	0,007	-0,0385	0,04125	0,00421	0,064471
Item27	-0,15	0,149	-0,0145	-0,04775	0,03025	-0,01176	0,063669
Item28	-0,211	0,123	-0,027	-0,06425	0,01225	-0,02933	0,061298
Item29	-0,123	0,191	0,0515	0,0115	0,08525	0,04768	0,058465
Item30	-0,14	0,143	0,011	-0,02675	0,0505	0,00849	0,058995
Item31	-0,13	0,12	-0,011	-0,0585	0,02825	-0,01454	0,057034
Item32	-0,24	0,108	-0,046	-0,09425	-0,00875	-0,05146	0,065757
Item33	-0,153	0,173	0,006	-0,035	0,07825	0,0209	0,07617
Item34	-0,233	0,119	-0,0365	-0,083	0,005	-0,04036	0,067588
Item35	-0,142	0,157	-0,024	-0,0615	0,0215	-0,01539	0,062631
Item36	-0,149	0,155	0,02	-0,0225	0,0555	0,01786	0,062047
Item37	-0,066	0,23	0,065	0,006	0,1055	0,05913	0,066553
Item38	-0,184	0,057	-0,06	-0,10425	-0,01875	-0,05999	0,054369
Item39	-0,129	0,129	0,0205	-0,01575	0,05275	0,01652	0,054075
Item40	-0,2	0,153	0,007	-0,03625	0,04725	0,00141	0,069486
Item41	-0,117	0,132	0,0255	-0,00225	0,0515	0,02568	0,044572
Item42	-0,179	0,168	-0,002	-0,036	0,032	0,00208	0,067146
Item43	-0,161	0,121	-0,025	-0,05325	0,01275	-0,02065	0,058776
Item44	-0,195	0,121	-0,0105	-0,04075	0,0365	-0,00848	0,059449
Item45	-0,176	0,098	-0,013	-0,04625	0,01675	-0,01636	0,050072
Item46	-0,118	0,126	0,0185	-0,014	0,0435	0,01885	0,047086
Item47	-0,194	0,143	0,005	-0,0455	0,04575	0,00159	0,060195
Item48	-0,215	0,127	-0,0405	-0,06975	-0,0025	-0,04061	0,058645
Item49	-0,123	0,165	0,002	-0,03625	0,03925	0,00158	0,057554
Item50	-0,146	0,176	0,01	-0,02525	0,05	0,01331	0,063943
Item51	-0,084	0,14	0,0215	-0,006	0,04475	0,02	0,042181
Item52	-0,223	0,083	-0,0795	-0,11025	-0,0355	-0,07325	0,061696
Item53	-0,104	0,106	0,0015	-0,044	0,035	-0,00382	0,048589
Item54	-0,119	0,129	0,0035	-0,03725	0,03525	0,00065	0,047105
Item55	-0,132	0,131	-0,018	-0,05225	0,01725	-0,01799	0,049192
Item56	-0,104	0,097	-0,012	-0,045	0,01825	-0,00984	0,048191
Item57	-0,078	0,102	0,015	-0,009	0,04925	0,01581	0,039055
Item58	-0,157	0,037	-0,067	-0,105	-0,027	-0,06492	0,04987
Item59	-0,33	0,13	0,0145	-0,0165	0,0425	0,01123	0,058258
Item60	-0,089	0,094	0,007	-0,02375	0,024	0,00301	0,039695

Tableau 7b. (suite) Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour la différence de difficulté standardisée modifiée (N = 100 échantillons).

Item	500 sujets						
	Minimum	Maximum	Médiane	1er quartile	3e quartile	Moyenne	Écart-type
Item1	-0,075	0,134	0,023	-0,0065	0,05125	0,02298	0,04393
Item2	-0,174	0,101	-0,04	-0,067	-0,01625	-0,03981	0,04541
Item3	-0,098	0,094	0,0135	-0,00625	0,03825	0,01535	0,035059
Item4	-0,077	0,145	0,0525	0,0275	0,07775	0,05107	0,041701
Item5	-0,059	0,14	0,0445	0,01	0,0735	0,04093	0,042335
Item6	-0,165	0,048	-0,0465	-0,079	-0,0195	-0,0477	0,045398
Item7	-0,148	0,033	-0,0645	-0,08725	-0,03975	-0,06308	0,04021
Item8	-0,153	0,089	-0,0205	-0,05725	0,0065	-0,02243	0,044366
Item9	-0,009	0,161	0,0685	0,0365	0,10025	0,07008	0,041903
Item10	-0,109	0,187	0,0415	0,0155	0,076	0,0456	0,049597
Item11	-0,12	0,059	-0,011	-0,0475	0,01025	-0,01979	0,04236
Item12	-0,079	0,148	0,0275	-0,011	0,065	0,02513	0,047421
Item13	-0,127	0,092	-0,021	-0,0565	0,0095	-0,0226	0,046752
Item14	-0,036	0,161	0,0755	0,05175	0,11025	0,07747	0,041097
Item15	-0,148	0,085	-0,038	-0,07175	-0,0115	-0,04076	0,04631
Item16	-0,051	0,127	0,0475	0,0155	0,0775	0,04451	0,042979
Item17	-0,22	0,015	-0,0855	-0,1155	-0,0555	-0,08426	0,048735
Item18	-0,039	0,152	0,0425	0,01275	0,074	0,04335	0,040287
Item19	-0,002	0,222	0,108	0,087	0,13	0,10734	0,038644
Item20	-0,08	0,099	0,021	-0,01825	0,04725	0,01498	0,042303
Item21	-0,075	0,182	0,0265	-0,00425	0,0655	0,03002	0,049068
Item22	-0,066	0,112	0,035	0,00075	0,06575	0,03278	0,042531
Item23	-0,192	0,029	-0,074	-0,105	-0,0515	-0,07885	0,0446
Item24	-0,088	0,123	0,0055	-0,02225	0,02825	0,00537	0,037356
Item25	-0,091	0,109	0,0145	-0,02025	0,052	0,01457	0,045385
Item26	-0,102	0,128	0,0095	-0,038	0,03825	0,00274	0,047987
Item27	-0,116	0,141	-0,0135	-0,048	0,01375	-0,01566	0,046994
Item28	-0,119	0,087	-0,0245	-0,052	0,00025	-0,02213	0,039265
Item29	-0,06	0,33	0,055	0,03575	0,08225	0,05661	0,045218
Item30	-0,101	0,1	0	-0,021	0,02025	-0,00125	0,03591
Item31	-0,093	0,1	-0,0155	-0,04175	0,01025	-0,01507	0,038255
Item32	-0,332	0,089	-0,0525	-0,09025	-0,011	-0,05093	0,066506
Item33	-0,101	0,18	0,0155	-0,01425	0,045	0,01436	0,047731
Item34	-0,118	0,053	-0,0405	-0,06	-0,01575	-0,0399	0,035887
Item35	-0,099	0,081	-0,0135	-0,03825	0,014	-0,01363	0,03801
Item36	-0,116	0,126	0,0195	-0,01325	0,04125	0,01567	0,04308
Item37	-0,108	0,201	0,0395	0,02175	0,0765	0,04433	0,046671
Item38	-0,153	0,02	-0,0685	-0,0925	-0,042	-0,0663	0,034997
Item39	-0,092	0,121	0,0235	-0,01425	0,05825	0,01944	0,046212
Item40	-0,086	0,1	0,0155	-0,01	0,041	0,01436	0,037168
Item41	-0,04	0,119	0,0315	0,009	0,06225	0,03455	0,035689
Item42	-0,088	0,082	0,009	-0,02125	0,03425	0,00496	0,037855
Item43	-0,112	0,068	-0,0175	-0,038	0,00825	-0,01745	0,036625
Item44	-0,121	0,07	-0,016	-0,04475	0,01225	-0,01759	0,037846
Item45	-0,092	0,077	-0,01	-0,03325	0,01125	-0,01092	0,033264
Item46	-0,09	0,102	0,0195	-0,00025	0,04825	0,02185	0,034088
Item47	-0,118	0,084	-0,015	-0,041	0,01325	-0,01494	0,041967
Item48	-0,129	0,044	-0,04	-0,0665	-0,00675	-0,03746	0,038663
Item49	-0,094	0,093	-0,0005	-0,026	0,02525	-0,00012	0,039612
Item50	-0,128	0,126	0,006	-0,021	0,0425	0,01168	0,04529
Item51	-0,061	0,085	0,019	0,0015	0,0395	0,0194	0,030049
Item52	-0,194	0,058	-0,066	-0,09775	-0,04275	-0,06931	0,046965
Item53	-0,121	0,079	-0,0115	-0,039	0,01825	-0,01094	0,038387
Item54	-0,114	0,061	-0,004	-0,02525	0,01225	-0,00604	0,029358
Item55	-0,132	0,053	-0,0245	-0,05225	0,0015	-0,02617	0,036449
Item56	-0,085	0,076	-0,0125	-0,03625	0,009	-0,01243	0,032722
Item57	-0,056	0,077	0,015	-0,0005	0,034	0,01726	0,025428
Item58	-0,14	0,015	-0,0635	-0,083	-0,04975	-0,06391	0,030178
Item59	-0,066	0,083	0,008	-0,0145	0,03	0,00833	0,028601
Item60	-0,075	0,056	-0,001	-0,01825	0,01725	-0,0005	0,026755

Tableau 7b. (suite) Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour la différence de difficulté standardisée modifiée (N = 100 échantillons).

Item	1000 sujets						
	Minimum	Maximum	Médiane	1er quartile	3e quartile	Moyenne	Écart-type
Item1	-0,18	0,083	0,016	0,00075	0,034	0,01496	0,034633
Item2	-0,11	0,038	-0,0405	-0,055	-0,01675	-0,0374	0,030306
Item3	-0,041	0,088	0,014	-0,006	0,0335	0,01476	0,027343
Item4	-0,015	0,116	0,0545	0,03275	0,07125	0,05202	0,029259
Item5	-0,048	0,126	0,0295	0,008	0,05525	0,03328	0,031451
Item6	-0,123	0,021	-0,055	-0,065	-0,029	-0,04898	0,028564
Item7	-0,139	0,04	-0,056	-0,0795	-0,03775	-0,05853	0,032252
Item8	-0,098	0,055	-0,0215	-0,036	-0,001	-0,02071	0,031319
Item9	-0,008	0,128	0,058	0,03475	0,08	0,05769	0,032099
Item10	-0,036	0,131	0,046	0,024	0,0725	0,04735	0,035437
Item11	-0,098	0,05	-0,0255	-0,04575	0,00125	-0,02094	0,034129
Item12	-0,045	0,101	0,0255	-0,001	0,045	0,02195	0,029006
Item13	-0,114	0,07	-0,028	-0,04325	-0,006	-0,02595	0,035936
Item14	0,009	0,148	0,075	0,0555	0,095	0,0756	0,029411
Item15	-0,119	0,053	-0,0455	-0,0635	-0,02	-0,04233	0,031703
Item16	-0,041	0,107	0,045	0,0225	0,06725	0,04465	0,029169
Item17	-0,167	0,011	-0,088	-0,11125	-0,06625	-0,08748	0,034414
Item18	-0,001	0,113	0,048	0,024	0,058	0,04472	0,02509
Item19	0,027	0,214	0,105	0,0925	0,128	0,10665	0,029707
Item20	-0,048	0,096	0,0155	-0,00925	0,035	0,0167	0,030716
Item21	-0,049	0,127	0,0355	0,01475	0,06	0,03759	0,033368
Item22	-0,051	0,099	0,0385	0,016	0,057	0,03392	0,031522
Item23	-0,159	0,013	-0,079	-0,09925	-0,0565	-0,07844	0,032435
Item24	-0,062	0,075	0,003	-0,01525	0,01725	0,00171	0,025947
Item25	-0,066	0,095	0,011	-0,01025	0,0235	0,00893	0,03021
Item26	-0,084	0,078	-0,006	-0,026	0,02	-0,00409	0,033479
Item27	-0,078	0,056	-0,0165	-0,02925	0,00275	-0,01313	0,02596
Item28	-0,088	0,055	-0,0205	-0,0385	0,00225	-0,01824	0,02812
Item29	-0,003	0,115	0,0575	0,043	0,07425	0,05811	0,02616
Item30	-0,066	0,081	0,002	-0,01425	0,019	0,00189	0,026643
Item31	-0,079	0,069	-0,0185	-0,03325	0,001	-0,01673	0,026582
Item32	-0,153	0,039	-0,0485	-0,0715	-0,03075	-0,05043	0,03466
Item33	-0,048	0,074	0,0165	-0,00725	0,03375	0,01461	0,029097
Item34	-0,12	0,03	-0,0435	-0,06175	-0,023	-0,04343	0,029775
Item35	-0,117	0,074	-0,0025	-0,02325	0,01125	-0,00502	0,032133
Item36	-0,065	0,096	0,0145	-0,00925	0,031	0,01066	0,029797
Item37	-0,022	0,144	0,046	0,02075	0,0705	0,04708	0,033115
Item38	-0,129	-0,013	-0,0645	-0,07825	-0,047	-0,06416	0,025951
Item39	-0,024	0,088	0,0205	0,00675	0,051	0,02647	0,027322
Item40	-0,33	0,085	0,0115	-0,00925	0,027	0,00675	0,044382
Item41	-0,029	0,1	0,0285	0,01175	0,042	0,02855	0,024724
Item42	-0,059	0,054	0,0065	-0,01525	0,02625	0,00535	0,026995
Item43	-0,071	0,04	-0,012	-0,03425	0,002	-0,01517	0,025337
Item44	-0,089	0,067	-0,01	-0,03125	0,011	-0,00848	0,029974
Item45	-0,076	0,048	-0,0145	-0,028	0,00725	-0,01148	0,025268
Item46	-0,038	0,066	0,019	0,00475	0,03425	0,01795	0,02253
Item47	-0,058	0,061	-0,003	-0,026	0,01425	-0,00531	0,026855
Item48	-0,11	0,078	-0,0315	-0,051	-0,0185	-0,0345	0,031153
Item49	-0,069	0,066	-0,008	-0,02225	0,0115	-0,00598	0,026781
Item50	-0,07	0,102	0,017	-0,00725	0,037	0,01306	0,030601
Item51	-0,03	0,07	0,0145	0,003	0,028	0,01513	0,022084
Item52	-0,144	0,014	-0,067	-0,08925	-0,04875	-0,06858	0,031871
Item53	-0,079	0,054	-0,0075	-0,0275	0,008	-0,00983	0,028582
Item54	-0,054	0,052	-0,003	-0,01825	0,008	-0,0039	0,020909
Item55	-0,077	0,05	-0,021	-0,04	0,002	-0,01699	0,026922
Item56	-0,083	0,054	-0,007	-0,02925	0,00925	-0,01008	0,025715
Item57	-0,022	0,063	0,017	0,007	0,031	0,01751	0,017127
Item58	-0,116	0,01	-0,065	-0,077	-0,05075	-0,06392	0,021296
Item59	-0,035	0,05	0,013	-0,002	0,02325	0,01197	0,018387
Item60	-0,036	0,039	0,0005	-0,01025	0,00925	4E-05	0,016743

Tableau 7b. (suite) Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour la différence de difficulté standardisée modifiée (N = 100 échantillons).

Item	2000 sujets						
	Minimum	Maximum	Médiane	1er quartile	3e quartile	Moyenne	Écart-type
Item1	-0,032	0,06	0,014	-0,003	0,02725	0,01333	0,021139
Item2	-0,105	0,012	-0,04	-0,05525	-0,02375	-0,03988	0,023583
Item3	-0,038	0,071	0,0195	0,008	0,034	0,02037	0,018914
Item4	-0,001	0,098	0,053	0,037	0,0685	0,05298	0,020363
Item5	-0,02	0,102	0,03	0,0125	0,04625	0,03071	0,023896
Item6	-0,088	0,001	-0,044	-0,06	-0,03	-0,04433	0,019943
Item7	-0,103	0,057	-0,06	-0,07525	-0,04575	-0,05884	0,023541
Item8	-0,081	0,025	-0,0245	-0,04325	-0,008	-0,02537	0,024619
Item9	0,013	0,12	0,0555	0,04375	0,074	0,05787	0,020424
Item10	-0,007	0,113	0,0455	0,029	0,06025	0,04462	0,022509
Item11	-0,075	0,052	-0,022	-0,037	-0,007	-0,02019	0,022377
Item12	-0,035	0,063	0,02	0,011	0,031	0,02112	0,018343
Item13	-0,088	0,037	-0,026	-0,04325	-0,01175	-0,02718	0,022946
Item14	0,027	0,117	0,0765	0,06475	0,08825	0,0761	0,017799
Item15	-0,088	0,003	-0,0445	-0,055	-0,02875	-0,04242	0,020008
Item16	-0,001	0,107	0,042	0,0275	0,05525	0,04195	0,021434
Item17	-0,133	0	-0,0865	-0,104	-0,067	-0,0839	0,025719
Item18	-0,003	0,097	0,05	0,03475	0,062	0,04838	0,020672
Item19	0,058	0,15	0,1045	0,09175	0,117	0,10435	0,017568
Item20	-0,049	0,088	0,013	-0,005	0,02225	0,01142	0,023622
Item21	-0,039	0,099	0,03	0,013	0,042	0,02886	0,024215
Item22	-0,026	0,069	0,0355	0,02	0,051	0,03439	0,020947
Item23	-0,131	0,075	-0,0815	-0,095	-0,06275	-0,07728	0,028347
Item24	-0,06	0,042	-0,004	-0,018	0,01	-0,00367	0,019558
Item25	-0,038	0,071	0,011	-0,00125	0,02225	0,01087	0,019982
Item26	-0,047	0,053	-0,005	-0,014	0,009	-0,00215	0,019848
Item27	-0,08	0,036	-0,014	-0,027	0,00025	-0,01328	0,02005
Item28	-0,061	0,043	-0,0205	-0,032	-0,00575	-0,01965	0,019048
Item29	0	0,107	0,053	0,042	0,06725	0,05413	0,018663
Item30	-0,049	0,066	0,0085	-0,01225	0,01925	0,00454	0,022279
Item31	-0,052	0,026	-0,012	-0,027	-0,001	-0,01323	0,016338
Item32	-0,108	0,05	-0,049	-0,06625	-0,03775	-0,05047	0,024671
Item33	-0,035	0,067	0,019	0,003	0,03325	0,01786	0,020712
Item34	-0,093	0,067	-0,041	-0,053	-0,02875	-0,04015	0,024211
Item35	-0,057	0,041	-0,005	-0,018	0,009	-0,00473	0,021027
Item36	-0,069	0,058	0,007	-0,004	0,019	0,00724	0,019528
Item37	-0,002	0,105	0,0515	0,032	0,065	0,05059	0,021302
Item38	-0,116	-0,006	-0,0615	-0,073	-0,05075	-0,06223	0,019276
Item39	-0,039	0,071	0,0215	0,00775	0,03925	0,02285	0,021938
Item40	-0,051	0,059	0,0015	-0,012	0,01625	0,00224	0,020815
Item41	-0,017	0,08	0,03	0,016	0,04	0,02826	0,017392
Item42	-0,042	0,057	0,004	-0,01	0,019	0,00473	0,019912
Item43	-0,054	0,039	-0,0165	-0,0255	0,004	-0,01149	0,021111
Item44	-0,05	0,044	-0,009	-0,025	0,004	-0,01001	0,019593
Item45	-0,039	0,022	-0,011	-0,024	0,00225	-0,01104	0,016296
Item46	-0,018	0,075	0,02	0,006	0,029	0,01835	0,016415
Item47	-0,038	0,051	0,002	-0,013	0,01225	0,00038	0,018047
Item48	-0,081	0,016	-0,033	-0,0475	-0,021	-0,03483	0,020217
Item49	-0,053	0,05	-0,0015	-0,012	0,013	0,00053	0,019617
Item50	-0,035	0,058	0,0105	-0,003	0,023	0,01078	0,01876
Item51	-0,015	0,047	0,0175	0,00975	0,02625	0,01808	0,013197
Item52	-0,133	-0,021	-0,065	-0,0795	-0,05175	-0,06698	0,020658
Item53	-0,055	0,036	-0,014	-0,02825	-0,0045	-0,01432	0,018999
Item54	-0,032	0,047	-0,0015	-0,01125	0,01025	-0,00021	0,016258
Item55	-0,054	0,029	-0,0155	-0,027	-0,00275	-0,01436	0,016936
Item56	-0,055	0,04	-0,012	-0,02425	0,002	-0,01078	0,018368
Item57	-0,012	0,053	0,021	0,012	0,029	0,02118	0,012335
Item58	-0,106	-0,019	-0,062	-0,07325	-0,0495	-0,06165	0,017545
Item59	-0,016	0,057	0,009	-0,001	0,02	0,00944	0,01495
Item60	-0,035	0,032	0,002	-0,006	0,012	0,0022	0,013772

Tableau 7c. Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour le khi carré de Mantel-Haenszel (N = 100 échantillons)

Item	250 sujets					500 sujets				
	Minimum	Maximun	Médiane	1er quartile	3e quartile	Minimum	Maximum	Médiane	1er quartile	3e quartile
Item1	0	8,98	0,185	0,04	0,8275	0	10,71	0,29	0,0575	1,47
Item2	0	11,92	0,275	0,02	1,175	0	15,81	0,93	0,1775	2,275
Item3	0	5,27	0,175	0,02	0,8125	0	4,07	0,295	0,0275	0,7525
Item4	0	9,65	0,74	0,1	2,0325	0	10,23	1,46	0,5	3,1875
Item5	0	8,24	0,335	0,04	1,255	0	7,47	0,575	0,09	1,79
Item6	0	5,43	0,575	0,0375	1,695	0	15,52	1,185	0,3575	3,5225
Item7	0	13,01	0,92	0,22	2,57	0	13,21	1,695	0,53	3,5125
Item8	0	9,4	0,295	0,03	1,355	0	7,45	0,285	0,07	1,2925
Item9	0	8,27	0,725	0,225	2,005	0	11,15	2,04	0,6725	5,285
Item10	0	6,42	0,335	0,02	1,0975	0	19,87	0,785	0,17	2,4525
Item11	0	7,12	0,35	0,0275	0,84	0	7,05	0,19	0,0275	1,1625
Item12	0	5,19	0,335	0,045	1,085	0	8,09	0,725	0,155	2,0375
Item13	0	10,5	0,6	0,07	1,46	0	8,79	0,48	0,0825	1,9975
Item14	0	11,9	1,52	0,46	4,1	0	15,16	3,09	1,525	7,13
Item15	0	11,56	0,455	0,0275	1,9825	0	8,39	0,63	0,1475	2,16
Item16	0	7,98	0,64	0,1975	1,6925	0	11,09	1,245	0,3375	2,79
Item17	0	8,83	1,59	0,465	4,585	0	21,36	3,36	1,6775	7,59
Item18	0	7,84	0,51	0,08	1,3575	0	10,16	0,78	0,0875	1,6925
Item19	0	17,71	2,57	0,8625	4,915	0,03	25,99	7,23	3,7975	10,2675
Item20	0	7,84	0,18	0,02	0,7375	0	5,08	0,21	0,02	1,05
Item21	0	8,84	0,425	0,0975	1,245	0	14,18	0,565	0,08	2,27
Item22	0	6,42	0,35	0,04	1,545	0	8,08	0,795	0,1575	2,175
Item23	0	7,13	0,915	0,1825	2,6025	0	15,23	2,115	0,85	4,445
Item24	0	8,59	0,165	0,02	1,0125	0	4,84	0,285	0,05	0,875
Item25	0	7,95	0,285	0,0375	1,005	0	5,69	0,325	0,0775	1,4725
Item26	0	8,05	0,195	0,02	0,7425	0	7,71	0,475	0,02	1,495
Item27	0	5,3	0,245	0,02	0,9675	0	9,25	0,275	0,05	0,89
Item28	0	12,06	0,39	0,03	1,305	0	5,97	0,46	0,0875	1,5875
Item29	0	8,14	0,88	0,15	2,6375	0	13,23	2,07	0,8825	4,24
Item30	0	5,7	0,285	0,03	1,0825	0	6,19	0,2	0,0275	0,71
Item31	0	5,04	0,28	0,02	0,7825	0	5,32	0,33	0,0275	1,2125
Item32	0	11,95	0,59	0,1375	1,815	0	20,14	1,87	0,355	4,89
Item33	0	7,56	0,33	0,08	1,665	0	8,22	0,545	0,075	1,515
Item34	0	6,81	0,61	0,1225	1,815	0	15,33	1,125	0,2075	3,28
Item35	0	6	0,4	0,06	1,06	0	6,45	0,32	0,0375	1,255
Item36	0	4,52	0,36	0,02	1,1925	0	8,88	0,445	0,1175	1,1775
Item37	0	10,47	0,825	0,0675	2,485	0	20,77	1,175	0,4775	3,145
Item38	0	11,74	0,74	0,1525	2,6975	0	18,6	2,975	1,33	5,405
Item39	0	5,65	0,35	0,0375	1,2475	0	7,06	0,69	0,13	1,615
Item40	0	6,51	0,28	0,05	1,33	0	5,87	0,27	0,04	0,9125
Item41	0	8,82	0,31	0,03	1,2	0	11,22	0,97	0,225	2,635
Item42	0	7,25	0,235	0,02	1,125	0	3,4	0,315	0,05	1,055
Item43	0	8,97	0,485	0,0675	1,2975	0	8,54	0,315	0,06	1,1125
Item44	0	6,89	0,31	0,06	0,8175	0	5,07	0,375	0,06	1,185
Item45	0	9,15	0,23	0,04	0,8175	0	6,82	0,36	0,05	1,215
Item46	0	8,47	0,255	0,03	1,125	0	10,97	0,485	0,08	1,8175
Item47	0	7,32	0,245	0,02	0,8375	0	6,25	0,345	0,0575	0,955
Item48	0	10,22	0,375	0,0575	1,5725	0	10,38	1,06	0,17	2,4675
Item49	0	7,14	0,38	0,05	0,985	0	5,76	0,39	0,0475	1,105
Item50	0	6,43	0,305	0,02	0,9125	0	8,98	0,41	0,06	1,1725
Item51	0	6	0,26	0,04	1,0575	0	6,53	0,59	0,1225	1,9525
Item52	0	12,9	1,115	0,2575	2,5675	0	17,3	2,41	0,775	5,0525
Item53	0	7,87	0,25	0,02	0,87	0	9,7	0,31	0,05	1,4675
Item54	0	5,69	0,32	0,06	0,91	0	6	0,175	0,02	0,81
Item55	0	9,15	0,265	0,02	0,815	0	10,77	0,595	0,17	1,7525
Item56	0	4,11	0,275	0,0375	1,135	0	7,73	0,465	0,045	1,3525
Item57	0	6,3	0,25	0,03	0,915	0	8,78	0,65	0,19	1,8225
Item58	0	10,56	1,22	0,27	3,5425	0,01	15,16	3,86	2,1125	6,2775
Item59	0	7,93	0,28	0,0375	1,5425	0	7,58	0,31	0,05	1,17
Item60	0	6,92	0,16	0,02	0,8625	0	4,9	0,19	0,0175	0,9025

Tableau 7c. (suite) Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour le khi carré de Mantel-Haenszel (N = 100 échantillons)

Item	1000 sujets					2000 sujets				
	Minimum	Maximum	Médiane	1et quartile	3 e quartile	Minimum	Maximum	Médiane	1er quartile	3e quartile
Item1	0	8,66	0,6	0,155	1,465	0	11,72	0,615	0,1175	1,6575
Item2	0	11,48	1,155	0,3425	3,3875	0	23,33	3,11	1,3075	6,1325
Item3	0	4,95	0,28	0,0575	1,3125	0	6,43	0,32	0,06	1,06
Item4	0	16,23	3,345	1,44	5,84	0	22,77	7,23	3,795	11,0675
Item5	0	13,09	0,765	0,18	2,4625	0	18,09	1,155	0,3325	3,71
Item6	0	12,38	3,03	1,045	5,035	0	15,64	4,36	1,835	7,3925
Item7	0,04	17,57	2,595	1,0225	5,6625	0,09	24,34	7,01	4,62	10,565
Item8	0	10,02	0,705	0,2375	1,75	0	13,21	1,64	0,41	3,1825
Item9	0	17,63	3,545	1,485	6,9025	0,08	27,68	7,435	4,7925	10,69
Item10	0	20,65	1,68	0,3925	4,3575	0	21,29	3,185	1,3675	5,94
Item11	0	8,07	0,865	0,155	2,0425	0	11,13	1,265	0,2425	2,4825
Item12	0	14,34	0,93	0,1575	2,2625	0	12,55	1,52	0,535	3,07
Item13	0	10,93	1,275	0,5725	3,0225	0	16,16	2,095	0,76	4,8525
Item14	0,14	31,09	7,515	4,115	11,5525	4,38	41,87	16,86	12,3825	21,845
Item15	0	9,29	1,53	0,21	3,8775	0	18,2	2,91	1,035	5,1525
Item16	0	12,52	2,52	0,8375	5,475	0	27,96	5,015	2,58	8,66
Item17	0,13	33,89	10,295	6,3175	14,745	0,12	48,66	19,915	12,8325	29,43
Item18	0	11,94	1,8	0,5975	3,7625	0,09	25,52	4,835	2,485	7,8125
Item19	1,86	52,05	14,195	10,8775	20,5875	10,86	65,97	29,805	24,6925	37,705
Item20	0	6,45	0,32	0,06	1,075	0	6,58	0,335	0,05	1,0325
Item21	0	13,64	1,8	0,5475	4,235	0	14,79	2,81	1,07	4,8925
Item22	0	8,83	1,84	0,565	3,6575	0	17,61	3,84	1,7975	6,5775
Item23	0,02	23,95	5,965	2,6325	8,2175	1,17	30,23	11,385	6,7825	16,3925
Item24	0	4,93	0,25	0,035	0,8625	0	8,97	0,42	0,06	0,9375
Item25	0	8,56	0,455	0,115	1,205	0	10,53	0,5	0,075	1,4975
Item26	0	5,94	0,37	0,0375	1,1775	0	7,67	0,345	0,085	1,3025
Item27	0	4,44	0,25	0,07	0,845	0	8,3	0,315	0,08	1,0125
Item28	0	7,8	0,74	0,15	1,915	0	7,99	1,46	0,3675	3,035
Item29	0,05	19,33	5,295	2,955	8,9075	0,02	31,84	9,29	5,725	13,8525
Item30	0	7,67	0,265	0,08	1,1425	0	9,66	0,84	0,1025	1,7975
Item31	0	9,23	0,645	0,1675	1,8575	0	8	0,79	0,1075	2,2425
Item32	0	27,25	3,505	1,6525	7,4625	0,85	33,45	8,515	4,9175	13,74
Item33	0	5,5	0,46	0,135	1,395	0	6,97	0,645	0,18	2,1625
Item34	0,01	19,4	3,415	1,2125	6,6325	0	26,58	6,99	3,765	11,2675
Item35	0	15,65	0,225	0,05	0,885	0	7,59	0,395	0,09	1,3525
Item36	0	11,59	0,43	0,1075	1,25	0	6,19	0,415	0,1	1,0625
Item37	0	23,56	3,315	0,93	6,055	1,05	25,58	7,145	4,1025	11,9775
Item38	0,1	28,69	4,85	2,7325	8,4125	1,8	35,71	10,925	7,885	16,89
Item39	0	7,84	0,755	0,08	2,2775	0	9,91	1,395	0,2775	3,1375
Item40	0	6,86	0,34	0,04	1,015	0	6,82	0,37	0,09	1,11
Item41	0	17,99	1,47	0,44	4,055	0,01	20,7	4,285	1,5225	8,2125
Item42	0	4,33	0,29	0,05	0,9375	0	5,82	0,365	0,0875	0,9725
Item43	0	4,79	0,44	0,1175	1,2725	0	10,87	0,865	0,235	2,1825
Item44	0	10,55	0,66	0,11	1,6925	0	7,84	0,565	0,09	1,355
Item45	0	10,85	0,52	0,145	1,5275	0	7,44	0,755	0,205	2,535
Item46	0	7,49	1,215	0,1975	2,2725	0,05	21,13	2,015	0,5475	3,945
Item47	0	8,26	0,47	0,0675	1,255	0	12,35	0,585	0,1675	1,4975
Item48	0	19,62	2,305	0,7725	4,1525	0,01	24,69	4,795	1,9825	8,4
Item49	0	7,93	0,37	0,0375	1,255	0	5,94	0,48	0,1	1,045
Item50	0	8,73	0,56	0,1375	1,45	0	8,26	0,395	0,085	1,4025
Item51	0	10,39	0,74	0,205	1,77	0	13,81	1,65	0,545	3,4525
Item52	0	24,9	4,815	2,54	8,91	1,12	39,3	10,855	7,3725	15,3875
Item53	0	8,58	0,59	0,18	1,875	0	10,58	1,27	0,3075	2,8125
Item54	0	7,66	0,36	0,0375	1,015	0	10,45	0,495	0,075	1,665
Item55	0	10,12	0,755	0,16	2,73	0	9,72	1,375	0,56	3,6025
Item56	0	10,47	0,575	0,0675	2,24	0	11,36	1,205	0,2775	3,25
Item57	0	15,78	1,235	0,22	3,42	0,01	18,12	4,29	1,9875	7,2375
Item58	0,31	24,13	8,6	5,63	11,8975	3,75	42,72	17,765	12,185	22,9625
Item59	0	8,19	0,405	0,1475	1,185	0	14,25	0,5	0,16	1,3825
Item60	0	3,77	0,235	0,03	0,8675	0	7,25	0,43	0,08	1,295

Tableau 7d. Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour le khi carré d'amlioration et la régression logistique (N = 100 échantillons)

Item	250 sujets					500 sujets				
	Minimum	Maximun	Médiane	1er quartile	3e quartile	Minimum	Maximum	Médiane	1er quartile	3e quartile
Item1	0,006	12,132	1,543	0,725	3,829	0,038	20,067	2,126	0,93375	4,05525
Item2	0,016	16,285	1,802	0,85325	3,85375	0	18,605	2,127	1,193	4,771
Item3	0,052	11,647	2,208	0,83425	4,1	0,014	9,194	1,515	0,623	3,037
Item4	0,023	14,739	2,1945	0,78025	3,84925	0,205	14,649	3,422	1,77	5,948
Item5	0,005	13,677	1,5925	0,89	3,676	0,003	9,262	2,1915	1,16825	3,50425
Item6	0,001	15,367	1,9465	0,75525	4,19625	0,011	20,74	2,5855	1,23925	5,291
Item7	0,032	14,378	2,6255	1,294	5,28475	0,078	17,356	3,6575	1,35475	5,90925
Item8	0,018	10,938	1,803	0,72925	3,59775	0,035	12,793	1,2125	0,52125	2,92175
Item9	0,066	13,648	2,2495	1,133	4,18025	0,112	14,229	3,4965	1,53425	6,51325
Item10	0,02	13,557	1,8575	0,799	3,82525	0,062	17,5	2,4005	1,02175	4,7275
Item11	0,007	13,373	1,215	0,6145	2,49425	0,006	8,939	1,578	0,48975	3,06125
Item12	0,001	9,585	1,311	0,50275	2,69025	0,012	13,668	1,9275	0,74475	3,82475
Item13	0,066	11,813	1,678	0,7585	3,1025	0,034	11,135	1,8225	0,70725	3,697
Item14	0,117	17,417	3,538	1,61625	6,10475	0,017	19,292	5,901	2,65825	9,40025
Item15	0,005	16,675	2,3305	0,69725	4,544	0,083	15,682	2,4225	1,135	4,619
Item16	0,017	9,795	1,696	0,63875	3,191	0,058	22,743	2,768	1,1025	4,432
Item17	0,021	15,35	3,893	2,147	7,56775	0,213	29,205	5,5635	3,1705	9,62325
Item18	0,066	9,932	1,455	0,7645	3,42175	0,001	17,846	2,2855	0,93125	3,82075
Item19	0,068	30,298	4,586	2,3645	7,5295	0,001	31,485	9,412	6,2335	13,435
Item20	0,013	12,796	1,581	0,5445	3,58575	0,037	17,903	1,8115	0,84625	3,28625
Item21	0,001	11,415	1,3595	0,61775	2,81425	0,014	17,012	1,997	0,8	4,1605
Item22	0,017	12,346	2,092	1,01925	3,36925	0,003	13,786	1,976	0,83525	4,2905
Item23	0,047	13,799	2,708	0,919	4,4175	0,072	18,515	4,504	2,20875	6,8045
Item24	0,028	9,746	1,5525	0,55975	2,9175	0,009	8,867	1,2225	0,45425	2,002
Item25	0,052	12,34	1,128	0,6315	2,90375	0,026	8,93	1,6215	0,54125	2,85025
Item26	0,024	10,422	1,3645	0,52675	2,46375	0,047	10,339	1,4955	0,80825	2,555
Item27	0,028	11,103	1,888	0,56725	3,28	0,03	11,436	1,668	0,674	3,14925
Item28	0,017	14,235	1,7095	0,6385	3,63325	0,007	9,381	1,9725	0,87	3,86825
Item29	0,102	32,939	2,4085	0,942	4,42225	0,003	19,359	4,5285	1,89475	6,96225
Item30	0,014	7,589	1,4105	0,645	2,54225	0,017	11,851	1,401	0,549	2,596
Item31	0,009	9,84	1,34	0,606	2,69125	0,007	10,526	1,716	0,84075	3,09975
Item32	0,007	16,465	2,2235	1,0965	4,2175	0,071	21,061	3,765	1,684	7,593
Item33	0,015	13,871	1,836	0,65225	4,537	0,043	26,528	2,0755	1,2125	3,77775
Item34	0,015	12,413	2,0785	0,81475	4,0555	0,028	17,391	2,9495	1,1805	5,469
Item35	0,011	13,722	1,794	0,74925	3,10275	0,021	12,799	1,3635	0,5935	3,26875
Item36	0,006	9,874	1,315	0,4695	2,89075	0,029	10,672	1,2965	0,562	3,03175
Item37	0,049	13,837	2,1255	0,91525	4,7935	0,139	20,846	2,537	0,86525	4,2215
Item38	0,014	14,731	2,673	1,39525	5,2385	0,062	20,627	5,25	2,64225	7,6355
Item39	0,008	6,103	1,41	0,6445	2,95725	0,043	8,314	1,606	0,84875	3,26075
Item40	0,008	12,604	1,6395	0,70525	3,2345	0,029	12,625	1,07	0,436	2,3145
Item41	0,068	13,466	1,543	0,54375	2,8565	0,011	12,709	2,158	1,017	4,563
Item42	0,017	13,239	1,556	0,52275	3,659	0,005	7,589	1,3775	0,59175	2,85025
Item43	0,048	11,844	1,619	0,6445	2,80925	0,009	8,203	1,5	0,68975	2,74525
Item44	0,014	9,09	1,0105	0,4325	2,28825	0,011	9,829	1,563	0,641	2,8285
Item45	0,003	14,745	1,2085	0,4165	2,14975	0,016	7,691	1,551	0,527	3,253
Item46	0,01	18,42	1,5305	0,5675	3,40075	0,055	15,077	1,853	0,72225	4,007
Item47	0,003	10,529	1,0045	0,41375	2,80125	0,008	9,246	1,6685	0,65175	3,2605
Item48	0,004	14,567	2,1825	0,98525	4,02025	0,055	14,267	3,0125	1,62825	5,10625
Item49	0,002	10,711	1,3845	0,5375	2,97825	0,001	7,254	1,201	0,54225	2,87125
Item50	0,001	11,289	1,1265	0,70625	2,7775	0,019	11,153	1,3475	0,5945	2,8125
Item51	0,02	15,505	1,5155	0,70925	2,70625	0,054	9,374	1,7935	0,75	3,88725
Item52	0,009	15,48	3,0125	1,417	4,888	0,015	19,484	4,64	2,378	7,52625
Item53	0,006	10,073	1,7995	0,8825	3,003	0,033	11,602	2,254	1,0835	4,67875
Item54	0,005	12,026	1,6005	0,78525	2,90075	0,004	11,542	1,068	0,42775	2,502
Item55	0,001	18,411	1,845	0,75875	3,3755	0,033	15,433	2,5185	1,20475	4,864
Item56	0,07	9,999	1,8685	1,00425	4,06	0	13,981	2,708	1,16525	5,266
Item57	0,032	11,259	1,647	0,6075	3,6745	0,099	19,655	2,252	0,91	4,322
Item58	0,027	17,252	4,156	1,41775	7,20075	0,642	19,075	6,62	4,0925	10,19325
Item59	0,002	16,804	1,488	0,72925	3,23825	0,012	11,284	1,427	0,85725	2,61175
Item60	0,082	14,527	1,599	0,66075	3,02975	0,091	13,845	2,138	0,85	3,99275

Tableau 7d. (suite) Distribution des indices de fonctionnement différentiel en fonction de la grandeur des échantillons pour le khi carré d'amlioration et la régression logistique (N = 100 échantillons)

Item	1000 sujets					2000 sujets				
	Minimum	Maximum	Médiane	1er quartile	3e quartile	Minimum	Maximum	Médiane	1er quartile	3e quartile
Item1	0,01	14,718	2,1815	0,65225	4,0785	0,112	25,114	2,384	1,18225	5,33525
Item2	0,095	18,167	3,7655	1,5605	6,6535	0,281	25,867	6,393	3,455	9,77225
Item3	0,152	16,721	2,2435	1,0595	3,70175	0,125	17,384	3,194	1,50325	6,2735
Item4	0,112	23,017	5,3785	2,938	9,03925	1,22	33,768	10,3565	6,03325	15,64975
Item5	0,005	14,821	1,93	0,85025	4,21925	0,033	20,568	2,972	1,0475	5,157
Item6	0,042	16,7	4,688	2,38225	7,07675	0,224	18,981	5,8775	3,7195	10,7315
Item7	0,089	17,094	4,35	2,1295	7,85425	1,805	24,538	8,785	5,3055	12,6015
Item8	0,112	9,55	1,667	0,9565	3,08	0	14,341	2,3685	1,06075	4,464
Item9	0,007	19,248	5,005	3,30525	9,2215	0,44	31,381	8,673	5,73625	12,98
Item10	0,027	20,621	3,205	1,5335	6,238	0,085	21,646	5,048	2,318	7,8425
Item11	0,001	14,442	2,2295	0,98375	3,884	0,01	17,043	2,502	1,00325	4,41675
Item12	0,004	17,224	1,9955	1,11125	3,29275	0,018	16,112	2,5075	1,2315	4,48825
Item13	0,055	15,755	3,1265	1,77275	5,496	0,422	20,746	4,6525	2,6155	7,84725
Item14	0,18	32,096	9,795	5,78175	14,8375	1,81	47,817	18,7845	14,3185	23,7685
Item15	0,025	14,78	4,1735	1,9515	6,4785	0,478	20,139	7,438	4,309	10,538
Item16	0,232	19,774	4,271	2,454	7,17325	0,025	31,197	6,791	4,3085	10,881
Item17	0,872	41,883	13,818	8,14225	18,02375	2,352	51,097	22,4745	14,97425	31,65625
Item18	0,094	13,449	3,342	1,76375	5,3085	0,331	30,12	6,684	4,1	9,966
Item19	1,82	59,197	16,253	13,2065	23,26125	10,326	70,605	33,319	27,299	41,12475
Item20	0,048	18,475	2,9745	1,36325	5,969	0,233	27,09	4,9995	2,0575	7,86175
Item21	0,005	15,436	2,794	1,077	6,0085	0,079	17,436	4,4445	2,10525	6,97275
Item22	0,05	11,633	3,19	1,23575	5,75	0,025	28,64	5,0945	2,91175	9,028
Item23	0,746	26,766	8,0675	4,03	10,83475	1,712	32,801	13,967	8,97275	19,215
Item24	0,023	9,226	1,3145	0,43875	2,4265	0,005	10,485	1,3085	0,611	2,229
Item25	0,018	9,601	1,3135	0,53	2,42525	0,004	12,885	1,4815	0,54125	2,96675
Item26	0,005	9,076	1,1865	0,559	2,9855	0,006	7,874	1,196	0,4285	2,84675
Item27	0,023	9,481	1,425	0,5815	3,03425	0,007	13,019	2,354	0,85875	4,2575
Item28	0,001	9,639	2,1705	1,00175	4,20925	0,005	13,329	3,1	1,53825	5,08125
Item29	0,617	22,26	7,4035	4,43425	11,57525	0,299	34,76	11,579	8,1755	17,436
Item30	0,002	8,856	0,9555	0,4	2,34175	0,056	12,26	1,673	0,7085	2,9635
Item31	0,064	10,618	1,7675	0,904	3,7885	0	12,745	2,281	1,0185	4,2505
Item32	0,378	28,209	6,922	4,1315	11,13925	0,818	37,22	11,8615	8,81075	17,99825
Item33	0,085	12,642	2,8275	1,2235	4,94775	0,132	25,236	4,8205	2,80025	8,16975
Item34	0,309	20,382	6,103	3,08525	9,26175	0,303	31,192	10,7105	6,5015	13,9975
Item35	0,022	14,498	1,5635	0,64875	2,89225	0,016	11,834	2,228	0,84675	3,46575
Item36	0,025	12,019	1,159	0,49225	2,5275	0,046	11,787	1,2355	0,6055	2,24625
Item37	0,094	24,233	4,945	2,05225	7,43625	1,679	25,039	8,4385	5,29875	12,15425
Item38	0,463	27,557	7,3075	4,03925	10,369	3,854	40,071	12,4905	9,92825	18,141
Item39	0,009	9,189	1,894	0,888	3,8725	0,014	106,181	2,074	0,9865	5,17325
Item40	0,059	8,692	1,3405	0,64375	2,7615	0,073	7,151	1,475	0,63375	2,6485
Item41	0,043	21,822	3,2925	1,5845	6,14875	0,149	25,688	5,795	2,81875	9,5325
Item42	0,014	7,555	1,8355	0,825	2,923	0,041	12,839	1,8475	0,9355	2,60675
Item43	0,013	7,902	1,45	0,56325	2,99225	0,05	14,977	2,6085	1,15875	4,0935
Item44	0,052	12,877	1,817	0,87025	3,22075	0,04	12,217	1,852	0,7385	3,28625
Item45	0,008	16,363	1,9675	0,78125	3,509	0,012	13,142	3,126	1,0305	4,9135
Item46	0,056	10,565	2,0085	1,05875	4,53075	0,232	29,814	3,6015	1,33875	5,616
Item47	0,008	11,773	1,527	0,723	2,8575	0,041	16,318	1,695	0,73025	3,27775
Item48	0,027	18,08	4,415	2,11475	7,65225	0,331	28,734	7,7355	3,864	11,05825
Item49	0,026	21,042	1,258	0,62225	2,75775	0,02	7,671	1,9275	0,64325	3,364
Item50	0,027	16,169	1,26	0,55	2,627	0,029	9,146	1,2525	0,62875	2,129
Item51	0,017	11,35	1,7535	0,8505	4,1435	0,048	16,255	2,956	1,4	5,16825
Item52	0,225	27,975	7,209	4,11875	11,78825	2,447	41,258	14,3745	9,124	18,24375
Item53	0,064	16,509	2,7725	1,5255	5,63675	0,46	17,255	4,3395	2,5255	7,748
Item54	0,003	7,965	1,245	0,6425	2,3745	0,004	13,578	1,507	0,54275	3,15575
Item55	0,072	14,718	4,222	2,001	6,17975	0,073	28,12	7,141	3,63575	10,6075
Item56	0,293	22,987	5,187	2,7995	8,68925	0,656	24,187	9,43	5,32725	14,095
Item57	0,001	14,876	2,7965	1,27425	5,48625	0,002	21,039	5,2795	2,947	7,9265
Item58	0,732	36,903	11,402	8,4095	16,33175	7,805	51,523	22,1615	16,05725	28,956
Item59	0,053	9,858	1,5785	0,57075	3,08525	0,008	17,771	1,9165	0,73325	3,04775
Item60	0,027	19,545	2,443	1,024	4,68175	0,198	22,881	4,31	2,13675	8,216

Tableau 8a. Répartition des coefficients de validité phi lorsque le critère de décision est le centiles C_5 ou les centiles $C_{2,5}$ et $C_{97,5}$ (n = 100 comparaisons).

Indices	**250 sujets**	**500 sujets**	**1000 sujets**	**2000 sujets**
Δ_{MH}				
-0,200	6			
-0,100	0	1		
0,000	23	7	2	
0,100	39	34	14	3
0,200	0	4	0	1
0,300	30	35	37	30
0,400	2	18	47	64
0,500		1		2
0,600				
0,700				
0,800				
DDSM				
-0,200	8			
-0,100	1	2		
0,000	23	7	1	
0,100	33	29	12	2
0,200	3	0	1	1
0,300	30	38	45	23
0,400	2	24	41	69
0,500				5
0,600				
0,700				
0,800				

Tableau 8a. (suite) Répartition des coefficients de validité phi lorsque le critère de décision est le centiles C_5 ou les centiles $C_{2,5}$ et C_{97} (n = 100 comparaisons).

Indices	250 sujets	500 sujets	1000 sujets	2000 sujets
X_{MH}				
-0,200	12			
-0,100	0			
0,000	25	16	2	
0,100	0	0	0	
0,200	46	41	25	9
0,300	17	43	73	91
0,400				
0,500				
0,600				
0,700				
0,800				
$X_{amél.}$				
-0,200	13	1		
-0,100	0	0		
0,000	38	26	6	1
0,100	0	0	0	0
0,200	37	38	28	15
0,300	12	35	66	84
0,400				
0,500				
0,600				
0,700				
0,800				

Tableau 8b. Répartition des coefficients de corrélation phi lorsque le critère de décision est le centile C_{90} ou les centiles C_5 et C_{95} (n = 100 comparaisons).

Indices	**250 sujets**	**500 sujets**	**1000 sujets**	**2000 sujets**
Δ_{MH}				
-0,200	2			
-0,100	15	2		
0,000	20	8	2	
0,100	36	28	7	2
0,200	3	1	2	0
0,300	19	38	30	23
0,400	4	17	40	32
0,500	1	6	18	41
0,600			1	2
0,700				
0,800				
DDSM				
-0,200	3			
-0,100	8	1		
0,000	21	7	2	
0,100	39	32	11	1
0,200	1	4	3	0
0,300	18	27	28	14
0,400	9	24	36	35
0,500	1	5	19	46
0,600			1	4
0,700				
0,800				

Tableau 8b. (suite) Répartition des coefficients de corrélation phi lorsque le critère de décision est le centile C_{90} ou les centiles C_5 et C_{95} (n = 100 comparaisons).

Indices	250 sujets	500 sujets	1000 sujets	2000 sujets
X_{MH}				
-0,200	3			
-0,100	11	1		
0,000	17	7		
0,100	31	17	7	1
0,200	1	1	0	0
0,300	25	43	21	7
0,400	11	22	46	31
0,500	1	9	26	61
0,600				
0,700				
0,800				
$X_{amél.}$				
-0,200	2			
-0,100	14	3		
0,000	28	8	1	
0,100	35	37	6	3
0,200	0	0	0	0
0,300	15	27	33	15
0,400	5	19	38	42
0,500	1	6	22	40
0,600				
0,700				
0,800				

Tableau 8c. Répartition des coefficients de corrélation phi lorsque le critère de décision est le centile C_{80} ou les centiles C_{10} et C_{90} (n = 100 comparaisons).

Indices	**250 sujets**	**500 sujets**	**1000 sujets**	**2000 sujets**
Δ_{MH}				
-0,200	4			
-0,100	4			
0,000	29	6		
0,100	26	16	5	
0,200	24	32	16	3
0,300	11	28	25	10
0,400	2	12	25	21
0,500		6	18	35
0,600			9	19
0,700			2	11
0,800				1
DDSM				
-0,200	4			
-0,100	2			
0,000	22	5		
0,100	29	18	3	
0,200	23	30	6	2
0,300	16	25	36	6
0,400	4	13	24	22
0,500		8	20	27
0,600		1	8	27
0,700			3	14
0,800				2

Tableau 8c. (suite) Répartition des coefficients de corrélation phi lorsque le critère de décision est le centile C_{80} ou les centiles C_{10} et C_{90} (n = 100 comparaisons).

Indices	250 sujets	500 sujets	1000 sujets	2000 sujets
X_{MH}				
-0,200	3			
-0,100	3	1		
0,000	30	7		
0,100	26	17	2	
0,200	19	1	16	
0,300	13	43	21	6
0,400	5	22	20	20
0,500	1	9	25	28
0,600			11	31
0,700			4	12
0,800			1	3
$X_{amél.}$				
-0,200	1			
-0,100	5			
0,000	36	11		
0,100	28	18	7	
0,200	21	34	13	2
0,300	6	21	29	14
0,400	2	11	25	31
0,500	1	5	16	29
0,600			7	18
0,700			3	3
0,800				3

Tableau 8d. Répartition des coefficients de validité phi lorsque le critère de décision est un critère fixé a priori (n = 100 comparaisons).

Indices	**250 sujets**	**500 sujets**	**1000 sujets**	**2000 sujets**
Δ_{MH}				
Moins de -0,200	5	1		
-0,100	4	0		
-0,200	17	4	1	
0,100	28	18	4	1
0,200	23	30	14	5
0,300	15	21	26	28
0,400	7	16	23	29
0,500	1	8	22	28
0,600		2	7	8
0,700			3	1
0,800				
DDSM				
Moins de -0,200	2			
-0,100	6			1
0,000	10		1	0
0,100	26	16	4	0
0,200	27	36	8	4
0,300	20	17	33	30
0,400	8	20	20	23
0,500	1	9	27	32
0,600		2	5	9
0,700			2	1
0,800				

Tableau 8d. (suite) Répartition des coefficients de validité phi lorsque le critère de décision est un critère fixé a priori (n = 100 comparaisons).

Indices	250 sujets	500 sujets	1000 sujets	2000 sujets
X_{MH}				
-0,200	7			
-0,100	4	2		
0,000	13	4		
0,100	24	16	1	
0,200	18	27	8	
0,300	26	24	19	2
0,400	6	14	21	12
0,500	1	9	32	36
0,600		3	13	33
0,700		1	5	15
0,800 et plus			1	2
$X_{amél.}$				
-0,200	6			
-0,100	5	1		
0,000	24	4		
0,100	25	24	2	
0,200	16	29	16	2
0,300	19	16	21	7
0,400	3	19	35	28
0,500	2	5	15	40
0,600		2	9	19
0,700			2	4
0,800				

Tableau 9a. Taux d'identification correcte d'après les centiles C95 ou C2,5 et C97,5 pour des échantillons de taille variée estimés à partir du modèle logarithmique

Nombre de sujets dans l'échantillon	Taux prédits pour les indices des méthodes à l'étude				
	Log(N)	Delta MH	DDSM	Khi car. MH	Khi car. amél.
250	5,521461	0,515536	0,5100761	0,5900928	0,519578
300	5,703782	0,551945	0,5485277	0,6261378	0,559743
400	5,991465	0,609395	0,6091999	0,6830125	0,62312
500	6,214608	0,653957	0,6562608	0,727128	0,672278
600	6,39693	0,690367	0,6947125	0,763173	0,712444
700	6,55108	0,721151	0,7272228	0,7936486	0,746403
800	6,684612	0,747817	0,7553846	0,8200477	0,77582
900	6,802395	0,771338	0,7802251	0,8433334	0,801768
1000	6,907755	0,792379	0,8024456	0,8641632	0,824978
1100	7,003065	0,811412	0,8225465	0,883006	0,845975
1200	7,090077	0,828788	0,8408972	0,9002082	0,865144
1300	7,17012	0,844773	0,8577782	0,9160326	0,882777
1400	7,244228	0,859572	0,8734076	0,9306838	0,899103
1500	7,31322	0,87335	0,8879582	0,9443237	0,914302
2000	7,600902	0,9308	0,9486303	1,0011984	0,977679

Tableau 9b. Taux d'identification correcte d'après les centiles C90 ou C5 et C95 pour des échantillons de taille variée estimés à partir du modèle logarithmique

Nombre de sujets dans l'échantillon	Taux prédits pour les indices des méthodes à l'étude				
	Log(N)	Delta MH	DDSM	Khi car. MH	Khi car. amél.
250	5,521461	0,475794	0,4919647	0,5195658	0,468604
300	5,703782	0,511	0,5264964	0,5562307	0,506016
400	5,991465	0,566552	0,5809834	0,6140835	0,565049
500	6,214608	0,609641	0,6232468	0,6589577	0,610838
600	6,39693	0,644847	0,6577785	0,6956226	0,64825
700	6,55108	0,674614	0,6869746	0,7266223	0,679882
800	6,684612	0,700399	0,7122655	0,7534754	0,707282
900	6,802395	0,723142	0,7345736	0,7771616	0,731451
1000	6,907755	0,743488	0,7545288	0,7983496	0,753071
1100	7,003065	0,761892	0,7725806	0,8175165	0,772629
1200	7,090077	0,778694	0,7890606	0,8350145	0,790484
1300	7,17012	0,79415	0,8042206	0,851111	0,806909
1400	7,244228	0,80846	0,8182567	0,8660142	0,822115
1500	7,31322	0,821783	0,8313239	0,8798886	0,836273
2000	7,600902	0,877334	0,8858109	0,9377415	0,895305

Tableau 9c. Taux d'identification correcte d'après les centiles C80 ou C10 et C90 pour des échantillons de taille variée estimés à partir du modèle logarithmique

Nombre de sujets dans l'échantillon	Taux prédits pour les indices des méthodes à l'étude				
	Log(N)	DeltaMH	DDSM	Khi car.MH	Khi car.amél.
250	5,521461	0,413091	0,4259994	0,422848	0,399838
300	5,703782	0,441479	0,4548791	0,454262	0,429046
400	5,991465	0,486271	0,500448	0,503829	0,475133
500	6,214608	0,521014	0,5357939	0,542277	0,51088
600	6,39693	0,549402	0,5646737	0,573691	0,540088
700	6,55108	0,573403	0,5890911	0,600251	0,564783
800	6,684612	0,594194	0,6102425	0,623259	0,586175
900	6,802395	0,612533	0,6288993	0,643553	0,605044
1000	6,907755	0,628937	0,6455884	0,661706	0,621922
1100	7,003065	0,643777	0,6606856	0,678128	0,637191
1200	7,090077	0,657325	0,6744682	0,69312	0,65113
1300	7,17012	0,669788	0,6871469	0,706912	0,663953
1400	7,244228	0,681326	0,6988856	0,71968	0,675825
1500	7,31322	0,692068	0,7098141	0,731568	0,686878
2000	7,600902	0,736861	0,7553829	0,781135	0,732965

Tableau 9d. Taux d'identification correcte d'après des critères fixés a priori pour des échantillons de taille variée estimés à partir du modèle logarithmique

Nombre de sujets dans l'échantillon	Taux prédits pour les indices des méthodes à l'étude				
	Log(N)	Delta MH	DDSM	Khi car. MH	Khi car. amél.
250	5,521461	0,371457	0,3879923	0,565811	0,487765
300	5,703782	0,416964	0,4329528	0,57015	0,493417
400	5,991465	0,48877	0,5038952	0,576997	0,502335
500	6,214608	0,544466	0,5589224	0,582308	0,509253
600	6,39693	0,589974	0,6038829	0,586647	0,514905
700	6,55108	0,62845	0,6418964	0,590316	0,519683
800	6,684612	0,661779	0,6748253	0,593494	0,523823
900	6,802395	0,691178	0,7038705	0,596297	0,527474
1000	6,907755	0,717476	0,7298525	0,598805	0,53074
1100	7,003065	0,741265	0,7533559	0,601073	0,533695
1200	7,090077	0,762983	0,7748129	0,603144	0,536392
1300	7,17012	0,782962	0,7945515	0,605049	0,538874
1400	7,244228	0,801459	0,8128265	0,606813	0,541171
1500	7,31322	0,81868	0,8298401	0,608455	0,54331
2000	7,600902	0,890485	0,9007825	0,615301	0,552228

Tableau 10a. Taux de vrais positifs d'après les centiles C95 ou C2,5 et C97,5 pour des échantillons de taille variée estimés à partir du modèle logarithmique

Nombre de sujets dans l'échantillon	Taux prédits pour les indices des méthodes à l'étude				
	Log(N)	Delta MH	DDSM	Khi car.MH	Khi car. amél.
250	5,521461	0,137663	0,139654	0,118102	0,104096
300	5,703782	0,147545	0,1498822	0,12534	0,112137
400	5,991465	0,163137	0,1660212	0,136761	0,124824
500	6,214608	0,175232	0,1785395	0,14562	0,134664
600	6,39693	0,185114	0,1887678	0,152858	0,142705
700	6,55108	0,193469	0,1974156	0,158978	0,149503
800	6,684612	0,200706	0,2049067	0,164279	0,155391
900	6,802395	0,20709	0,2115143	0,168955	0,160586
1000	6,907755	0,2128	0,2174251	0,173138	0,165232
1100	7,003065	0,217966	0,222772	0,176922	0,169435
1200	7,090077	0,222682	0,2276533	0,180376	0,173272
1300	7,17012	0,22702	0,2321437	0,183554	0,176802
1400	7,244228	0,231037	0,2363012	0,186496	0,18007
1500	7,31322	0,234777	0,2401717	0,189235	0,183113
2000	7,600902	0,250369	0,2563106	0,200656	0,1958

Tableau 10b. Taux de vrais positifs d'après les centiles C90 ou C5 ou C95 pour des échantillons de taille variée estimés à partir du modèle logarithmique

Nombre de sujets dans l'échantillon	Taux prédits pour les indices des méthodes à l'étude				
	Log(N)	Delta MH	DDSM	Khi car. MH	Khi car.améli
250	5,521461	0,192265	0,1995347	0,208717	0,186508
300	5,703782	0,206832	0,2138469	0,223303	0,20144
400	5,991465	0,229818	0,23643	0,246317	0,225001
500	6,214608	0,247647	0,2539467	0,264169	0,243276
600	6,39693	0,262215	0,268259	0,278754	0,258209
700	6,55108	0,274531	0,2803598	0,291086	0,270833
800	6,684612	0,2852	0,290842	0,301769	0,28177
900	6,802395	0,294611	0,300088	0,311192	0,291416
1000	6,907755	0,30303	0,3083588	0,31962	0,300045
1100	7,003065	0,310645	0,3158406	0,327245	0,307851
1200	7,090077	0,317597	0,322671	0,334206	0,314977
1300	7,17012	0,323993	0,3289544	0,34061	0,321533
1400	7,244228	0,329914	0,3347719	0,346538	0,327602
1500	7,31322	0,335426	0,3401878	0,352058	0,333253
2000	7,600902	0,358412	0,3627708	0,375072	0,356814

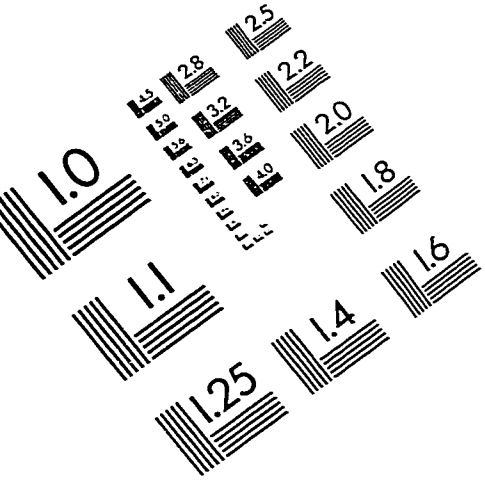
Tableau 10c. Taux de vrais positifs d'après les centiles C80 ou C10 et C90 pour des échantillons de taille variée estimés à partir du modèle logarithmique

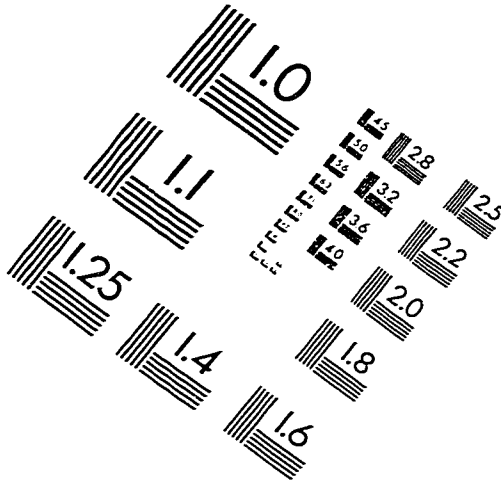
Nombre de sujets dans l'échantillon	Taux prédits pour les indices des méthodes à l'étude				
	Log(N)	DeltaMH	DDSM	Khi car.MH	Khi car.amél.
250	5,521461	0,332312	0,3447942	0,339101	0,319851
300	5,703782	0,355777	0,3685325	0,36417	0,343225
400	5,991465	0,392801	0,4059887	0,403726	0,380106
500	6,214608	0,42152	0,435042	0,434409	0,408713
600	6,39693	0,444985	0,4587802	0,459478	0,432086
700	6,55108	0,464824	0,4788507	0,480674	0,451848
800	6,684612	0,48201	0,4962364	0,499034	0,468967
900	6,802395	0,497168	0,5115718	0,515229	0,484067
1000	6,907755	0,510728	0,5252897	0,529716	0,497574
1100	7,003065	0,522995	0,5376991	0,542822	0,509793
1200	7,090077	0,534193	0,549028	0,554786	0,520948
1300	7,17012	0,544494	0,5594496	0,565791	0,531209
1400	7,244228	0,554032	0,5690984	0,575981	0,54071
1500	7,31322	0,562911	0,5780813	0,585468	0,549555
2000	7,600902	0,599936	0,6155375	0,625024	0,586436

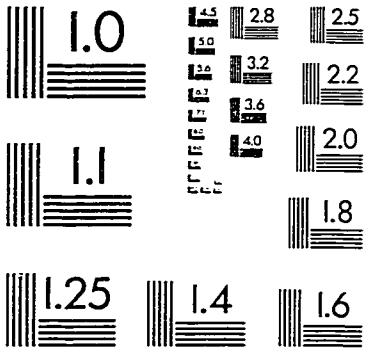
Tableau 10d. Taux de vrais positifs d'après des critères fixés a priori pour des échantillons de taille variée estimés à partir du modèle logarithmique

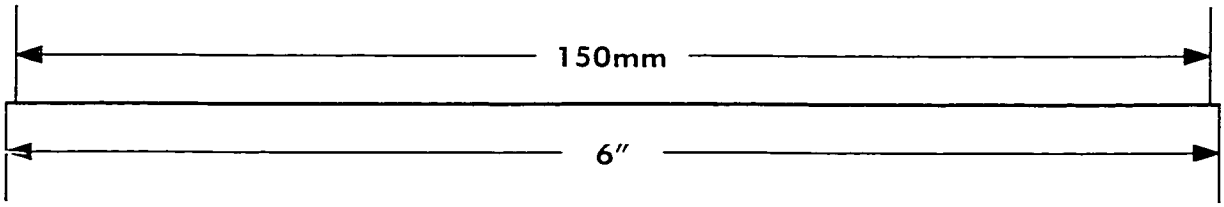
Nombre de sujets dans l'échantillon	Taux prédits pour les indices des méthodes à l'étude				
	Log(N)	Delta MH	DDSM	Khi car. MH	Khi car. amél.
250	5,521461	0,439564	0,4461302	0,139864	0,173927
300	5,703782	0,428643	0,4349174	0,200723	0,230228
400	5,991465	0,411411	0,4172249	0,296751	0,319064
500	6,214608	0,398045	0,4035016	0,371236	0,387971
600	6,39693	0,387124	0,3922888	0,432095	0,444272
700	6,55108	0,37789	0,3828086	0,483551	0,491874
800	6,684612	0,369892	0,3745964	0,528123	0,533108
900	6,802395	0,362837	0,3673527	0,567439	0,56948
1000	6,907755	0,356525	0,3608731	0,602609	0,602015
1100	7,003065	0,350816	0,3550115	0,634423	0,631447
1200	7,090077	0,345604	0,3496603	0,663468	0,658316
1300	7,17012	0,34081	0,3447376	0,690186	0,683033
1400	7,244228	0,336371	0,34018	0,714923	0,705917
1500	7,31322	0,332238	0,3359369	0,737953	0,727222
2000	7,600902	0,315006	0,3182445	0,833981	0,816059

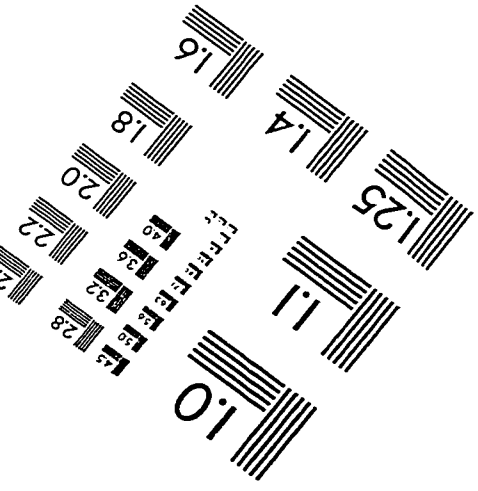
IMAGE EVALUATION TEST TARGET (QA-3)

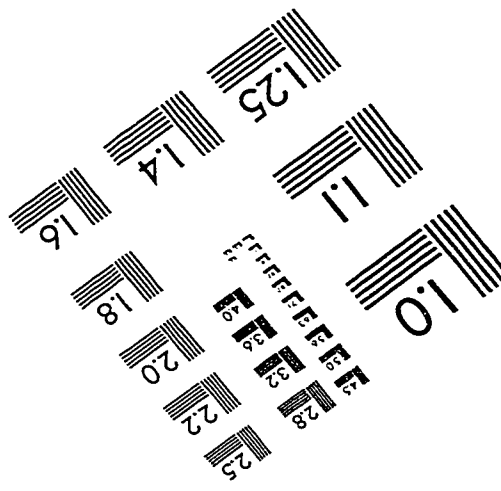












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989