MATHÉMATIQUES ET SCIENCES HUMAINES

ANDRÉ FLIELLER

Méthodes d'étude de l'adéquation au modèle logistique à un paramètre (modèle de Rasch)

Mathématiques et sciences humaines, tome 127 (1994), p. 19-47

<http://www.numdam.org/item?id=MSH_1994__127__19_0>

Math. Inf. Sci. hum. (32e année, n° 127, *1994*, p.19-47)

MÉTHODES D'ÉTUDE DE L'ADÉQUATION AU MODÈLE LOGISTIQUE A UN PARAMÈTRE (MODÈLE DE RASCH)

André FLIELLER*

RÉSUMÉ - *Le modèle de Rasch, ou modèle logistique de réponse à l'item à un paramètre, constitue une avancée méthodologique importante, mais l'étude de l'adéquation de données empiriques à ce modèle ne va pas sans problème. Après une brève présentation du modèle de Rasch, l'article discute certains problèmes généraux rencontrés dans les études d'adéquation. Vingt-quatre tests d'adéquation, graphiques et statistiques sont ensuite présentés et évalués. La nécessité d'employer plusieurs méthodes d'évaluation est soulignée dans la conclusion.*

SUMMARY- Methods of assessment of Rasch's model-data fit : A review.
Studying model data fit is a problem while employing Rasch model (one-parameter item response model). After a brief presentation of the Rasch model, this article discutes some general problems encountererd in the evaluation of item fit. Then 24 tests of item fit, both graphic and statistical, are presented and evaluated. The necessity of employing several methods of evaluation is enhanced in the conclusion.

Les modèles de réponse à l'item, dont le plus connu est le modèle de Rasch, constituent une avancée méthodologique importante, susceptible d'intéresser les chercheurs de diverses sciences sociales, notamment la psychologie, les sciences de l'éducation et la sociologie. Très utilisés aux États-Unis, au Canada et dans divers pays européens, tels que la Grande-Bretagne, les Pays-Bas ou le Danemark, ils demeurent largement ignorés des chercheurs français. Cette situation devrait évoluer favorablement sous l'effet de plusieurs facteurs, parmi lesquels l'intérêt porté par plusieurs équipes de recherche à la mesure adaptative, une des applications privilégiées de ces modèles, et, surtout, le développement des logiciels pour micro-ordinateurs, qui permettent de mettre en oeuvre ces modèles de façon beaucoup plus simple qu'il y a une dizaine d'années.

Cet article offre des repères méthodologiques au chercheur désireux d'utiliser le modèle de Rasch. Il est centré sur les méthodes permettant d'évaluer l'adéquation des données au modèle . Cette évaluation, qui joue un rôle central dans l'utilisation judicieuse du modèle, suscite en effet de nombreuses difficultés. L'information fournie par les manuels des logiciels étant trop sommaire, le chercheur se trouve renvoyé à une littérature spécialisée et éparse dont la lecture lui fait rapidement découvrir que chaque méthode pose des problèmes dont beaucoup n'ont pas de solution véritablement satisfaisante. D'autre part, les tests d'adéquation proposent des critères de décision que seule une expérience suffisante permet d'appliquer judicieusement. Appliqués aux mêmes données, ils conduisent parfois à des conclusions contradictoires (on en trouvera un exemple dans Reiser, 1989) qui laissent l'utilisateur désemparé. L'adéquation des données au modèle pose des problèmes particuliers dans le cas du modèle de Rasch. La simplicité de ce modèle a pour contrepartie des exigences très fortes qui ont peu de chances

*ADEPS (URA CNRS 1167), Université de Nancy II, C.O. n° 26, F-54015 Nancy Cedex

d'être satisfaites en totalité. En effet, comme le fait observer Hambleton (1989, p. 173), "the less restrictive the assumptions, the more likely a model fits the data to which it is applied". Lorsqu'on étudie une échelle à l'aide du modèle de Rasch, on est amené à constater dans la plupart des cas qu'elle n'y est pas conforme. On peut certes espérer que la suppression de certains items permettrait d'obtenir une meilleure adéquation. Mais il reste à déterminer lesquels et à décider si l'échelle obtenue est acceptable ou non, ce qui est souvent loin d'être évident.

Ce sont ces difficultés qui nous ont conduit à proposer la présente revue de question. Elle comporte trois parties. La première présente le modèle. Cette partie donne quelques détails sur l'estimation des paramètres en raison du lien qui unit cette question et celle de l'évaluation de l'adéquation. La deuxième partie examine les problèmes généraux que l'on rencontre lorsqu'on étudie l'adéquation des données au modèle. La troisième expose et évalue vingt-quatre tests d'adéquation.

1. LE MODÈLE DE RASCH

1.1. Objectif du modèle

Le modèle de Rasch (1960/1980) est une tentative pour rapprocher les sciences sociales des sciences physiques et biologiques. Selon l'auteur (1977), la différence essentielle entre ces deux catégories de sciences est le degré d'objectivité des mesures. Rasch entend «objectivité» dans un sens particulier qu'il convient de préciser. L'objectivité dont il est question ne concerne pas l'indépendance de l'observation par rapport à l'observateur, condition généralement satisfaite dans les sciences sociales. L'objectivité au sens de Rasch concerne l'indépendance de la mesure par rapport à l'instrument de mesure. Cette deuxième forme d'objectivité, appelée *objectivité spécifique* par Rasch, est atteinte dans les sciences physiques, mais ne l'est pas habituellement dans les sciences sociales. Une mesure physique, telle que la longueur d'un objet, est indépendante de l'instrument qui permet de l'obtenir, en ce sens qu'il n'est pas besoin de spécifier l'instrument utilisé (double-mètre, règle graduée, mètre de couturière, etc.). Une grande partie des mesures effectuées dans les sciences sociales ne peuvent au contraire être interprétées sans référence à l'instrument employé. Par exemple, la signification du score obtenu par un sujet dans un test lexical dépend de l'épreuve utilisée ; elle varie selon que les questions posées sont difficiles ou élémentaires. Pour remédier à cet inconvénient, on effectue habituellement une transformation du score brut destinée à situer la personne dans un groupe de référence. Mais ces mesures normatives dépendent étroitement du groupe de référence choisi, de sorte que le problème n'est pas réglé, mais seulement déplacé. La mesure des objets manque tout autant d'objectivité que celle des personnes. Comment par exemple caractériser le degré de difficulté d'un item ? La fréquence relative des réussites à l'item du test donne une indication, mais elle dépend à l'évidence du groupe sur laquelle elle est calculée. La mesure d'un item dépend donc des personnes auxquelles on l'applique et la mesure d'une personne dépend des items ou, dans le cas des mesures normatives, du groupe de référence choisi.

Le modèle de Rasch vise à surmonter ce double écueil. Il se propose de fournir des mesures objectives, c'est-à-dire des mesures de personnes indépendantes à la fois les unes des autres et des stimuli utilisés, et des mesures de stimuli indépendantes des personnes auxquelles on les applique. Avant d'exposer ce modèle, il nous paraît utile d'en préciser le champ d'application. Le modèle été élaboré dans le contexte de la mesure des compétences cognitives. Rasch (1977) explique que le problème de l'objectivité des mesures lui est apparu crucial quand il a été chargé d'évaluer les capacités de lecture de jeunes adultes. La terminologie utilisée («compétence» de l'individu, «difficulté» de l'item) porte la trace de cette origine. Cependant les problèmes mentionnés ci-dessus sont des problèmes généraux qui ne dépendent ni du type d'instrument utilisé (questionnaire, échelle d'évaluation, test...), ni du type de concept mesuré (attitude, utilité, aptitude...), ni du domaine empirique où la mesure est pratiquée. Le modèle de Rasch présente donc un intérêt général pour les sciences sociales. En s'en tenant à la littérature française -au demeurant très limitée- on peut trouver des exemples d'application à des questions

aussi diverses que la mesure de la pauvreté (Dickes, 1983), de la délinquance (Dickes & Hausman, 1983), des capacités de raisonnement (Flieller, 1989) ou de la dépression (Bonis, Féline, Lebeaux & Simon, 1994).

1.2. Le modèle

Le modèle de Rasch concerne les items dichotomiques où les deux réponses possibles, codées 1 et 0, sont ordonnées. Les données analysables à l'aide du modèle se présentent sous la forme d'une matrice (N x k) dont les cellules correspondent aux réponses de N sujets à k items dichotomiques.

Le modèle de Rasch est un modèle probabiliste de réponse à l'item : il exprime la relation entre la probabilité d'une réponse spécifiée (celle codée 1) et la mesure du sujet et de l'item dans une dimension inobservable, appelée dimension latente, dont l'item est un indicateur. Le modèle suppose que la probabilité de la réponse ne dépend que de deux facteurs, l'un concernant le sujet, l'autre l'item. Le premier facteur est la mesure du sujet dans la dimension latente, que l'on appelle *compétence du sujet* (θ), mais qu'il serait préférable d'appeler avec Andrich (1988, p. 25) *localisation du sujet* sur la dimension latente. Plus la compétence du sujet est élevée, plus la probabilité qu'il fournisse la réponse codée 1 est élevée. Le deuxième facteur de variation de la réponse est la mesure de l'item dans la dimension latente, appelée difficulté ou *localisation de l'item* (b) : plus elle est élevée, plus la probabilité que le sujet donne la réponse 1 est basse. La probabilité de la réponse à l'item est une fonction logistique de (θ - b), fonction choisie en raison de sa croissance monotone, de son intervalle de variation et de sa simplicité relative.

Le modèle s'écrit :

$$P_{si} = e^{(\theta-b)} / 1 + e^{(\theta-b)}, \qquad [1]$$

où P_{si} est la probabilité que le sujet s fournisse la réponse codée 1 à l'item i, θ est la localisation de s, et b la localisation de i. On déduit de [1] la valeur du rapport entre P_{si} et sa complémentaire Q_{si} :

$$P_{si} / Q_{si} = e^{(\theta-b)} \qquad [2]$$

qui constitue une autre écriture du modèle.

Pour un item donné, la courbe représentative de P_{si} est appelée *courbe caractéristique de l'item* (CCI). Le point d'inflexion de la courbe a pour coordonnées (b, 0,5). Il apparaît ainsi que la localisation de l'item correspond à la valeur de θ pour laquelle l'item apporte le maximum d'information.

1.3. Hypothèses du modèle

Le modèle implique deux hypothèses principales : l'*unidimensionnalité* et l'*indépendance locale.*

Selon l'hypothèse d'unidimentionnalité les items mesurent une même et unique dimension latente. Le concept d'unidimensionnalité appelle quelques remarques. En premier lieu l'unidimensionnalité est relative à un niveau d'observation, en l'occurrence l'item. Elle exprime l'existence d'une cohérence entre les réponses d'un sujet confronté à des items différents, cohérence que l'on explique par le concept mesuré. Mais comme le fait remarquer Whitely (1980), l'unidimensionnalité ainsi constatée n'est pas contradictoire avec l'existence de processus différents sous-jacents aux réponses fournies, qu'une observation plus fine permettrait de distinguer. En second lieu, l'unidimensionnalité des données n'est jamais

parfaite. Elle n'est, comme le souligne Reuchlin (1992), qu'une hypothèse vérifiée à un certain niveau d'approximation. L'hypothèse est acceptée quand les écarts des données par rapport au modèle peuvent être tenues pour aléatoires. En troisième lieu, un jugement d'unidimensionnalité dépend de l'opérationnalisation choisie, comme le fait bien apparaître la revue de Hattie (1985).

Selon la deuxième hypothèse du modèle, la probabilité d'un vecteur $(u_1, u_2,..., u_k)$ de réponses est égale au produit des k probabilités de réponse du sujet aux items :

$$P(\mathbf{u} \mid \theta) = P(u_1 \mid \theta)\, P(u_2 \mid \theta) \ldots P(u_k \mid \theta), \qquad [3]$$

où **u** est le vecteur de réponses et u_i la réponse du sujet à l'item i ($u_i = 1$ ou $u_i = 0$). La condition d'indépendance locale exige donc que la réponse d'un sujet à un item ne soit pas affectée par les réponses qu'il a données aux items antérieurs. C'est une hypothèse forte dont on ne se prive pas de discuter le réalisme (*v.g.* Goldstein, 1980) et qui n'est d'ailleurs que rarement testée. Elle constitue sans doute une des limites les plus importantes du modèle de Rasch ; mais cette réserve s'applique aux nombreux autres modèles de mesure qui supposent, eux aussi, l'indépendance locale. Si cette condition est satisfaite, la corrélation entre n'importe quelle paire d'items est nulle quand on tient θ constant ; la variable latente est donc la seule source de covariation des items, ce qui veut dire en d'autres termes que l'échelle est unidimensionnelle. Réciproquement, si l'échelle est unidimensionnelle, la variable latente constitue la seule source de covariation des items et la condition d'indépendance locale est satisfaite. Bien qu'il n'y ait pas unanimité sur ce point, on peut donc considérer que *l'indépendance locale est une opérationnalisation de l'unidimensionnalité* (Gustafsson, 1980 ; Hambleton, 1989 ; Lord & Novick, 1968).

1.4. Propriétés du modèle

L'échelle, qui mesure conjointement θ et b, est une *échelle d'intervalles* dont l'origine est indéterminée : comme on peut le vérifier facilement, P_{si} ne change pas si l'on ajoute ou retranche une constante à θ et à b. Il faut donc fixer l'origine. La solution la plus fréquente consiste à standardiser le paramètre b.

Une propriété spécifique du modèle de Rasch est que les items sont caractérisés uniquement par leur localisation. Pour cette raison le modèle de Rasch est fréquemment appelé *modèle à un paramètre* (sous-entendu «d'item» puisque le modèle comporte également un paramètre relatif au sujet). Cette caractéristique constitue une différence capitale entre le modèle de Rasch et d'autres modèles logistiques qui admettent que les items puissent varier également par leur discrimination, (c'est-à-dire la pente de leur tangente au point d'inflexion de la CCI, qui reflète le pouvoir discriminant de l'item : la variation de la probabilité de la réponse codée 1 résultant d'une variation de θ est d'autant plus forte que la pente est forte). Dans le modèle de Rasch, *tous les items ont la même discrimination.* Il s'ensuit que le modèle de Rasch est un *modèle hiérarchique* : $(\forall\, \theta)$ si $b_i > b_j$, alors $P_{si} > P_{sj}$. Le modèle de Rasch s'apparente ainsi au modèle de Guttman.

La propriété la plus remarquable du modèle est l'objectivité des mesures qu'il autorise. Cette objectivité se traduit par l'*invariance des paramètres* : si une échelle est raschienne, la mesure d'un sujet ne dépend pas du sous-ensemble d'items et, réciproquement, la localisation de l'item ne dépend pas du sous-échantillon de sujets. L'invariance de la localisation des sujets sur le trait latent a des applications intéressantes. Elle rend possible la comparaison de mesures obtenues dans une étude longitudinale avec des instruments différents (sous réserve bien sûr que leurs items forment une échelle raschienne). Elle facilite aussi grandement la mesure adaptative, qui consiste à sélectionner dans un pool d'items ceux les mieux adaptés au sujets, ceci permet d'obtenir des mesures plus précises avec moins d'items que dans la procédure classique, où les mêmes items sont donnés à tous les sujets. L'invariance de la localisation des items permet également de détecter les items biaisés, ce qui constitue une application intéressante pour les comparaisons de groupes (nationalité, classe sociale, sexe, etc.). On dit

qu'un item est biaisé quand P_{si} ne dépend pas seulement de la mesure de s dans la dimension latente, mais également du groupe auquel il appartient. Ce fonctionnement différentiel de l'item (*differential item functioning*) peut être détecté à l'aide du modèle de Rasch en comparant la valeur de b dans les deux groupes : une variation significative de b est indicatrice d'un biais. Insistons sur le fait que l'invariance des paramètres est une propriété du modèle ; pour que des données empiriques bénéficient de cette propriété, il faut qu'elles soient conformes au modèle.

1.5. Estimation des paramètres

Un dernier point à considérer dans cette présentation est l'estimation des paramètres b et θ

Un des avantages du modèle de Rasch par rapport aux autres modèles logistiques est que le nombre de paramètres à estimer y est plus faible. D'une part, il n'y a qu'un paramètre à estimer par item. D'autre part, on démontre que *le score r du sujet est une statistique suffisante* pour l'estimation de θ, de sorte que les sujets ayant le même score dans l'échelle ont la même mesure dans la dimension latente. Le nombre des paramètres à estimer passe ainsi de (k+N-1) à 2(k-1), N étant le nombre de sujets et k le nombre d'items[1].

La méthode d'estimation la plus courante est celle du maximum de vraisemblance. En supposant la condition d'indépendance locale satisfaite, la vraisemblance de la matrice des réponses est :

$$L(\mathbf{u}_1,\mathbf{u}_2,...,\mathbf{u}_n \mid \Theta, \mathbf{b}) = \Pi_n \Pi_k \mathbf{P}_{si}^{u} \mathbf{Q}_{si}^{1-u} \qquad [4]$$

où $\mathbf{u}_S$ est le vecteur des réponses du sujet s (*i.e.* une ligne de la matrice), Θ est le vecteur des N paramètres θ, **b** est le vecteur des k paramètres b, et où $\mathbf{P}_{si}^{u} \mathbf{Q}_{si}^{1-u}$ est la probabilité de la réponse u du sujet s à l'item i (probabilité égale à P_{si} si u = 1 et à Q_{si} si u = 0).

La méthode consistant à estimer les paramètres en maximisant la fonction [4] est appelée *méthode du maximum de vraisemblance conjoint* (JML : *joint maximum likelihood*). Cette méthode pose de sérieux problèmes. La question de savoir si le maximum est absolu ou seulement local est débattue. La non consistance des estimateurs obtenus par le JML est, en revanche, reconnue par tous les auteurs. Le JML conduit en effet à estimer simultanément des paramètres structuraux (b_i), dont le nombre est fixe, et des paramètres incidents (θ_S), dont le nombre augmente en même temps que le nombre de sujets. Or on montre que les estimateurs des paramètres structuraux ne peuvent être consistants quand ils sont estimés en même temps que des paramètres incidents. Andersen (1973, a) en a fourni la démonstration pour le modèle de Rasch. Le JML est cependant utilisé dans des programmes répandus, tel BICAL.[2]

Comme la non consistance des estimateurs des paramètres structuraux est due à la présence des paramètres incidents, on peut résoudre le problème en ne faisant pas intervenir la localisation des sujets dans l'estimation de la localisation des items. C'est l'objectif de la *méthode du maximum de vraisemblance conditionnel* (CML : *conditional maximum likelihood*) proposée par Andersen (1972). Le score r du sujet étant une statistique suffisante pour θ il est possible d'exprimer la fonction de vraisemblance en fonction de r. La probabilité pour un sujet s localisé en θ de fournir un vecteur **u** de k réponses sachant que son score total est r est égale au rapport entre la probabilité conditionnelle de **u** par rapport à θ et de la probabilité conditionnelle de r par rapport à θ :

$$P(\mathbf{u} \mid r, \mathbf{b}) = P(\mathbf{u} \mid \theta, \mathbf{b}) / P(r \mid \theta, \mathbf{b}). \qquad [5]$$

---

[1] Comme il est impossible d'estimer la compétence des sujets dont le score est nul ou parfait, il y a (k-1) scores possibles et donc 2(k-1) paramètres à estimer.

[2] Wright, auteur de BICAL, a montré que le JML fournissait des estimations satisfaisantes lorsqu'on introduisait un facteur de correction (Wright & Douglas, 1977).

On établit que

$$P(\mathbf{u} \mid r, \mathbf{b}) = = exp\ (-\Sigma_k \mathbf{u}_i \mathbf{b}_i) / \gamma_r, \qquad [6]$$

où u_i est la réponse à l'item i, b_i est la localisation de i et où γ_r, appelé *fonction symétrique élémentaire d'ordre r*, est défini ainsi :

$$\gamma_r = \Sigma_r\ exp(-\Sigma_k\ \mathbf{u}_i\ \mathbf{b}_i), \qquad [7]$$

Σ_r symbolisant la somme étendue aux C_k^r vecteurs de réponses possibles permettant d'obtenir le score r. Ainsi, *en conditionnant P (**u**) par rapport à r, on élimine le paramètre θ.* La fonction de vraisemblance pour N sujets se déduit directement de [6] :

$$L(\mathbf{u}_1, \mathbf{u}_2,...,\mathbf{u}_N \mid \mathbf{r}, \mathbf{b}) = exp(-\Sigma_N \Sigma_k\ \mathbf{u}_i\ \mathbf{b}_i) / \Sigma_N \Sigma_k\ \mathbf{u}_i\ \mathbf{b}_i / \Pi_N \gamma_r\ , \quad [8]$$

où **r** est le vecteur des N scores r des sujets. Cette fonction de vraisemblance ne fait pas intervenir les paramètres incidents, de sorte que les estimateurs des paramètres structuraux (localisation des items) sont consistants. Le CML présente l'avantage supplémentaire sur le JML de comporter des tests d'adéquation aux propriétés statistiques connues (Gustafsson[3] 1980, b). Cependant le calcul des fonctions symétriques pose certains problèmes techniques. La fonction symétrique d'ordre r est une somme comportant C_k^r produits de r termes. Dans une échelle de 60 items par exemple la fonction symétrique d'ordre 30 comporte plus de $1{,}18.10^{17}$ produits de 30 termes. On est donc obligé de plafonner le nombre d'items. Les premiers programmes (tel celui de Wright & Douglas, 1977) ne permettaient pas d'analyser des échelles de plus de 15 items. L'algorithme de Gustafsson (1980, a), mis en oeuvre dans le programme PML, a permis de repousser cette limite ; dans la version actuelle de PML pour micro-ordinateur elle est fixée à 60 items. Ces problèmes de calcul expliquent pourquoi la localisation des sujets, qu'on pourrait en principe estimer en conditionnant P(**u**) par rapport au score de l'item, est estimée en donnant aux paramètres b_i les valeurs précédemment obtenues par le CML.

L'estimation de la localisation des items par le CML est propre au modèle de Rasch. Dans les autres modèles logistiques, cette méthode d'estimation est impossible, parce que r n'est pas une statistique suffisante. On a donc recherché un autre moyen de s'affranchir des paramètres incidents. Bock et Aitkin (1981) ont proposé une méthode appelée méthode du *maximum de vraisemblance marginal* (MML : *marginal maximum likelihood*). Elle suppose de connaître la fonction de densité de θ. En appelant $g(\theta)$ cette fonction et $P(\mathbf{u} \mid \theta)$ la probabilité du vecteur $(u_1,u_2,...,u_k)$ pour un sujet localisé en θ la probabilité dite «marginale» de **u** est :

$$P(\mathbf{u}) = \int_{-\infty}^{+\infty} P(\mathbf{u} \mid \theta)\ g(\theta)\ d(\theta). \qquad [9]$$

Si la matrice de réponses des N sujets comporte v vecteurs distincts représentés f fois chacun, sa vraisemblance est :

$$L(\mathbf{u}_1, \mathbf{u}_2,...,\mathbf{u}_N) = \Pi_v\ P(\mathbf{u})^f \qquad [10]$$

Cette fonction ne fait pas intervenir les paramètres incidents. On trouvera dans l'article de Bock et Aitkin des précisions techniques sur la résolution des équations de vraisemblance. Thissen (1982) montre sur un exemple que les estimateurs obtenus par le MML et le CML sont proches. Signalons toutefois que le MML exige des effectifs plus importants et des temps de calculs plus longs que le CML. Le MML est la méthode d'estimation utilisée par BILOG (Mislevy & Bock, 1990), logiciel très employé et fiable si l'on se réfère aux simulations de Yen (1987).

---

[3] L'auteur considère même que la consistance présente un intérêt secondaire, dans la mesure où les estimateurs obtenus par le JML sont, *après correction*, proches de ceux obtenus par le CML.

Un des inconvénients du maximum de vraisemblance est qu'il ne permet pas d'estimer la localisation des items et des sujets à score parfait ou nul. Ces items auraient une difficulté infiniment grande ou petite et ces sujets auraient une compétence infiniment grande ou petite, ce qui n'a empiriquement pas de sens. Pour éviter ce problème on supprime tout simplement les items et sujets à score parfait ou nul. Mais cette procédure a des conséquences pratiques fâcheuses ; elle empêche par exemple le calcul de la compétence moyenne des sujets de l'échantillon initial. C'est une des raisons qui ont poussé Swaminathan et Gifford (1982) à élaborer une *procédure d'estimation bayésienne*. Pour le modèle de Rasch, l'intérêt d'une procédure bayésienne concerne surtout l'estimation de la localisation des sujets[4] ; c'est d'ailleurs la procédure par défaut de BILOG. Supposons que la localisation des items ait été estimée, par le MML par exemple, et que l'on veuille estimer la localisation des sujets. L'estimation bayésienne consiste à supposer que la distribution de θ est connue *a priori* et à exploiter cette information dans la procédure d'estimation de θ. Le théorème de Bayes permet d'établir que la densité *a posteriori* de θ est égale au produit de la fonction de vraisemblance de la matrice des réponses et de la densité *a priori* des N valeurs de θ :

$$f(\theta_1, \theta_2, \theta_N \mid \mathbf{u}_1,\mathbf{u}_2,..., \mathbf{u}_N) = L(\mathbf{u}_1,\mathbf{u}_2,..., \mathbf{u}_N \mid \theta_1,\theta_2,...,\theta_N) \times (\theta_1,\theta_2,...,\theta_N), \quad [11]$$

formule que l'on peut écrire de façon plus concise ainsi :

$$f(\Theta \mid \mathbf{U}) = L(\mathbf{U} \mid \theta)\, \Pi_N\, f(\theta_s) \quad [12]$$

où Θ désigne le vecteur des localisations des sujets, et **U** la matrice des réponses observées. La fonction de vraisemblance $L(\mathbf{U} \mid \theta)$ est donnée par la formule [4] ; la seule différence, mais importante, est que $L(\mathbf{U})$ n'est pas conditionnée par **b** puisque l'on connaît les valeurs des b_i. On suppose en général que θ suit la loi normale réduite ; $f(\theta_s)$ est donc la densité de θ_s dans la loi normale réduite. Les estimateurs bayésiens de θ sont les valeurs qui maximisent $f(\Theta \mid \mathbf{U})$.

Le lecteur intéressé par le modèle de Rasch gagnera à compléter son information sur les diverses méthodes d'estimation des paramètres. Il pourra se reporter aux publications originales mentionnées et à l'excellente revue de question de Baker (1987). Pour des informations complémentaires sur le modèle de Rasch, on peut consulter les manuels de Andrich (1988) et de Wright et Stone (1979), ainsi que l'ouvrage de Hambleton et Swaminathan (1985) consacré aux modèles de réponse à l'item.

2. PROBLÈMES GÉNÉRAUX RENCONTRÉS DANS LES ÉTUDES D'ADÉQUATION

Nous avons regroupé dans cette partie un certain nombre de problèmes généraux posés par l'évaluation de l'adéquation des données au modèle. Deux de ces problèmes sont liés aux tests d'adéquation eux-mêmes. On se demande s'ils sont suffisamment sensibles à la multidimensionnalité éventuelle de l'échelle étudiée (§ 2.1) et on montre que l'utilisation, souvent nécessaire, d'une distribution asymptotique nuit à l'objectivité de l'évaluation de l'adéquation (§ 2.2). Les trois problèmes suivants dépendent davantage des choix opérés par le chercheur. Le premier est le niveau d'analyse : l'échelle ou l'item ; les avantages et inconvénients de chaque niveau sont discutés. Le deuxième problème concerne la suppression des items non conformes, pratique courante chez les chercheurs utilisant le modèle de Rasch mais cependant discutable. Le troisième problème, plus technique, concerne le critère à utiliser pour effectuer une partition de l'échantillon lorsque le test l'exige, ce qui est fréquent. L'habitude est d'effectuer une partition selon le score total du sujet, mais elle est vivement critiquée par certains auteurs. Le sixième paragraphe attire l'attention sur les contraintes que les logiciels imposent au chercheur. Le dernier paragraphe signale les effets de l'estimation des paramètres sur la mesure de l'adéquation, effets insuffisamment pris en compte en général.

---

[4] Il en va différemment du modèle à deux paramètres où le principal intérêt de l'estimation bayésienne est de contraindre le paramètre de discrimination à rester dans des limites acceptables.

2.1. Sensibilité des tests à la multidimensionnalité

Plusieurs auteurs ont mis en évidence l'insensibilité possible des tests d'adéquation usuels à la violation de la condition d'unidimensionnalité définie ci-dessus (§ 1.3). Goldstein (1980) obtient un bon indice global d'adéquation pour un test de mathématiques alors que deux facteurs correspondant l'un aux items d'algèbre, et l'autre aux items de géométrie, sont trouvés par ailleurs. Gustafsson et Lindblad (1978, cités par Gustafsson, 1980) obtiennent des résultats similaires avec un questionnaire dont le caractère multidimensionnel est établi dans une étude préalable. Van den Wollenberg (1982, b, p. 126) énonce un ensemble de conditions qui, réunies, sont suffisantes pour que la bidimensionnalité d'une échelle ne puisse être détectée à l'aide des tests d'adéquation courants : "When two Rasch homogeneous tests have equal item parameters vectors and the traits follow the same distribution and when furthermore the sample is partitioned on the basis of the raw score, then the concatenation of the two tests behaves Rasch homogeneously in the sense that the item parameter estimates are, within chance limits, equal over samples". Sous les conditions énoncées, deux échelles raschiennes toutes deux, mais mesurant deux dimensions latentes différentes forment donc, lorsqu'elles sont réunies, une échelle conforme au modèle de Rasch pour tous les tests d'adéquation impliquant une partition de l'échantillon sur la base du score total des sujets. L'auteur montre à l'aide de données simulées que les prédictions théoriques sont vérifiées empiriquement. Le théorème de van den Wollenberg, dont on trouvera une interprétation géométrique dans l'article cité, est plus important théoriquement que pratiquement, car les conditions stipulées ont peu de chances d'être satisfaites. Mais il prouve que des tests solidement fondés et jugés par ailleurs trop puissants, tels le test d'Andersen, peuvent donner l'illusion d'une excellente adéquation au modèle, alors que l'échelle analysée est composée d'items se rapportant à des dimensions conceptuellement distinctes. Les tests d'adéquation les plus solides sont donc faillibles. Notons que le théorème se borne à définir un ensemble de conditions suffisantes pour que les tests échouent. Selon l'auteur il est possible que d'autres conditions aboutissent au même résultat.

Il est donc recommandé de ne pas se fier aveuglément aux test d'adéquation et de vérifier l'unidimensionnalité par d'autres méthodes. Des indications techniques peuvent être trouvées dans Hambleton (1989) et dans Hattie (1985). Un inconvénient des méthodes préconisées par ces auteurs est qu'elles sont souvent lourdes à mettre en oeuvre. On peut donc réserver leur emploi aux cas où des raisons théoriques amènent à douter de l'unidimensionnalité de l'échelle. Un contrôle est alors indispensable.

Les réserves faites dans ce paragraphe invitent à un certaine prudence mais ne doivent pas faire oublier que dans de nombreux cas la violation de l'unidimensionnalité est détectée par les tests d'adéquation. L'article de van den Wollenberg comporte d'ailleurs des simulations qui le montrent.

2.2. La distribution des statistiques d'adéquation

La distribution des statistiques d'adéquation n'est connue le plus souvent que de façon asymptotique. Ceci entraîne un problème bien connu (*v.g.* Hambleton & Swaminathan, 1985 ; Traub & Lam 1985 ; Rogers & Hattie, 1987) : pour estimer correctement la probabilité de la statistique d'adéquation il faut disposer d'un effectif important de sujets, mais un tel effectif risque de rendre trop puissant le test utilisé. On peut remarquer avec Hambleton (1989) qu'il existe toutefois deux cas privilégiés où l'interprétation des résultats obtenus est assez claire : celui où l'adéquation est bonne, bien que l'effectif des sujets soit élevé ($N > 1000$), et celui où l'adéquation est mauvaise, bien que l'effectif soit faible ($N < 300$). Lorsque l'effectif est important, ce qui est souhaitable, la solution habituelle au problème de la surpuissance consiste à abaisser le seuil de signification (0,001 est un seuil courant). Mais cette solution empirique comporte beaucoup d'inconvénients, faute de norme fiable et consensuelle sur le seuil à adopter en fonction du test utilisé, de la taille de l'échantillon et du nombre d'items. Le chercheur se voit exposé au risque de choisir le seuil en fonction des résultats qu'il souhaiterait obtenir. Le

débutant manque totalement de repères pour choisir un seuil convenable. L'évaluation externe d'une recherche est difficile en l'absence d'un critère objectif pour juger les décisions prises par l'auteur. Il serait à notre avis fort utile de déterminer au moyen de simulations les seuils les mieux adaptés à divers cas de figure.

Il arrive parfois que l'on connaisse la distribution exacte de la statistique d'adéquation, mais que l'on préfère l'approcher par une autre distribution qui simplifie les calculs. Cette approximation n'est légitime que si l'effectif des sujets est suffisamment important. Ce n'est généralement pas le cas pour les tests ultra-analytiques (cf. § 2.3 ci-dessous), puisque l'analyse porte sur des sous-échantillons à faibles effectifs. La distribution exacte doit être préférée à la distribution asymptotique dans ce type de test (Molenaar, 1983).

2.3. Le choix du niveau d'analyse

L'adéquation des données au modèle peut être étudiée soit au niveau de l'échelle, soit au niveau de l'item. Lorsqu'on se place au niveau de l'item, on peut considérer soit l'échantillon des sujets dans sa totalité, soit des sous-échantillons correspondant à des niveaux de compétence différents. Il existe donc trois niveaux d'analyse possibles. La littérature ne retient ordinairement que les deux premiers et oppose les tests globaux, portant sur la totalité de l'échelle, aux tests analytiques, portant sur chacun des items ; nous suggérons d'appeler «ultra-analytiques» les méthodes (graphiques, tests statistiques) permettant d'étudier l'adéquation des données au modèle par item et par groupe de compétence. Remarquons que la distinction entre tests globaux et tests analytiques concerne davantage les niveaux d'analyse que les tests eux-mêmes, car la plupart des tests analytiques permettent d'obtenir par sommation une mesure de l'adéquation globale.

Chaque niveau d'analyse présente ses avantages et ses inconvénients et aucun ne devrait être privilégié systématiquement. L'avantage des tests globaux est leur caractère synthétique. En revanche, ils sont souvent trop puissants et beaucoup d'auteurs renoncent à les employer. Ce point de vue se laisse discuter. Les conditions dans lesquelles les tests globaux sont trop puissants ne sont pas bien connues et demanderaient à être préciser par des études de simulation. En attendant, il faut se garder d'invoquer trop facilement l'argument de la surpuissance de ces tests pour expliquer la non conformité des données analysées. D'autre part le problème de puissance peut être évité en utilisant les tests globaux de manière descriptive. On peut ainsi étudier comment évolue l'adéquation de l'échelle quand elle subit des modifications. Les tests analytiques sont très employés parce qu'ils permettent d'identifier les items non conformes. Mais leur utilisation est problématique lorsqu'elle est solidaire d'une pratique consistant à supprimer les items non conformes sur la seule base des tests d'adéquation, stratégie dont nous verrons les limites au prochain paragraphe. Quant aux tests ultra-analytiques, leur intérêt est de permettre un diagnostic approfondi de l'adéquation d'un item, dont des exemples sont donnés par Hambleton, Swaminathan et Rogers (1991) et par Molenaar (1983). Mais ils sont exposés à des artefacts (effectif insuffisant de certains groupes, multiplication des comparaisons augmentant fortement le risque d'erreurs de première espèce) et ne sont pas d'emploi commode.

2.4. L'élimination d'items non conformes.

Comme nous l'avons indiqué dans la première partie, les hypothèses du modèle de Rasch sont si fortes qu'elles laissent en général peu d'espoir d'obtenir d'emblée une échelle conforme au modèle. Si l'on a de bonnes raisons de penser que cette inadéquation ne tient pas à des raisons fondamentales (échelle multidimensionnelle, effet d'apprentissage introduisant un lien de dépendance entre les réponses aux items, discrimination des items trop variable, etc.), il paraît légitime d'écarter quelques items inadéquats et d'analyser à l'aide du même modèle la nouvelle échelle ainsi obtenue. Ce procédé est très usité par les praticiens du modèle de Rasch. Malheureusement beaucoup d'entre eux éliminent drastiquement de nombreux items jusqu'à l'obtention d'une échelle raschienne, sans avoir conscience des problèmes posés par cette procédure.

Cette pratique constitue une dérive par rapport aux idées Rasch (1960/1980), qui a cherché à construire un modèle confirmatoire. Le modèle n'est pas un instrument de sélection aveugle des items. Son utilisation normale suppose au contraire que le chercheur ait construit une échelle dont l'unidimensionnalité est théoriquement fondée et dont les items sont théoriquement hiérarchisés.

Ce rôle nécessaire de la théorie est souligné par Whitely (1980) à partir d'autres considérations. Si la validité d'une échelle tient au fait que les items qui la composent permettent d'appréhender plusieurs facettes d'un même concept (par ex. l'intelligence générale mesurée par un test composite), éliminer les items inadéquats n'a pas pour effet d'accroître la validité de l'instrument, mais au contraire de l'abaisser.

L'élimination mécanique des items est vivement déconseillée par Gustafsson (1980, b) qui donne plusieurs arguments contre cette pratique. Le plus fondamental est qu'elle est illogique à ses yeux : "the basic requirement of the Rasch model is that the items shall be homogeneous, so what is tested is, in fact, wether the items fit with each other, not wether they fit the model" (p.229). Gustafsson estime que l'adéquation obtenue après l'élimination de nombreux items risque d'être illusoire. Rogers et Hattie (1987) souscrivent à cette opinion. En appliquant plusieurs tests d'adéquation à des données simulées ne satisfaisant pas certaines hypothèses du modèle, les auteurs constatent l'insensibilité de ces tests à certaines déviations et en concluent que "the exclusion of persons or items identified as misfitting by the statistics provide no assurance of fit of the remaining data" (p.56).

Un autre argument contre cette pratique est le phénomène de perturbation de la mesure des items adéquats par les items inadéquats (cf. § 2.5 ci-dessous), qui a pour conséquence le risque d'élimination prématurée d'items adéquats. Si l'on tient à éliminer des items, on ne saurait trop suivre les recommandations de Molenaar (1990) : éliminer les items pas à pas en se fondant, non pas sur un indicateur unique, mais sur la convergence des informations fournies par plusieurs tests.

Lorsque les items qui composent une échelle se laissent regrouper en sous-ensembles théoriquement homogènes (identité de contenu, ou de forme, ou de processus hypothétiquement en jeu, etc.), une solution meilleure que l'élimination d'items consiste à tester l'homogénéité de chacune de ces sous-échelles, puis à tester l'hypothèse qu'elles mesurent la même dimension latente au moyen du test ML-PCC de Martin-Löf. Cette suggestion de Gustafsson (1980, b) nous semble très fondée. De plus, elle est facile à mettre en oeuvre avec le logiciel PML-PC.

2.5. La partition de l'échantillon selon le score total

De nombreux tests d'adéquation impliquent un fractionnement de l'échantillon des sujets en deux ou plusieurs sous-échantillons. Le plus souvent, la partition se fait en fonction du score total r des sujets, soit que le test repose sur ce mode de subdivision, soit que le chercheur préfère ce critère pour des raisons théoriques ou simplement pratiques (il est toujours disponible, même si l'on ne dispose d'aucune information externe sur les sujets).

Ce critère est discuté par certains auteurs (McDonald, 1982 ; Molenaar, 1983 et 1990). On lui reproche principalement ce que l'on pourrait appeler une contamination de la mesure d'adéquation des "bons" items (ceux qui forment une échelle raschienne acceptable) par les "mauvais" items (ceux qui mesurent une autre dimension latente ou dont la discrimination a une valeur très différente des items du premier groupe). La localisation des sujets est estimée sur la base de leur score total. Cette estimation fait donc intervenir non seulement les bons items mais également les mauvais. Supposons que i soit un bon item. Le calcul de P_{si} à partir du modèle utilise une estimation erronée de θ. Il s'ensuit que cette probabilité est inévitablement différente de la fréquence observée. La mesure de l'adéquation d'un bon item souffre ainsi de la présence de mauvais items. Hambleton (1969, cité par Hattie, 1985, p. 154) génère dix items raschiens

de mauvais items. Hambleton (1969, cité par Hattie, 1985, p. 154) génère dix items raschiens auxquels il en ajoute cinq mesurant une autre dimension latente. Il constate non seulement une inadéquation globale de l'échelle (ce qui est rassurant), mais également une mauvaise adéquation de la plupart des dix items raschiens. Une minorité d'items suffit donc à faire suspecter d'inadéquation une majorité d'items parfaitement conformes au modèle !

Face à ce problème Molenaar (1983) rejette catégoriquement le score total comme critère de subdivision de l'échantillon et préconise de le remplacer par un critère externe tel que le sexe des sujets. Nous ne sommes pas convaincu qu'un critère externe suffise à résoudre le problème. Un tel critère n'a d'intérêt que si l'on peut supposer des différences sensibles entre les sous-groupes considérés. Selon la distinction de Andrich (1988, p. 69), ces différences peuvent être de degré ou de nature. Le premier cas correspond à une différence de localisation moyenne entre les groupes. Subdiviser le test suivant un critère externe ou un critère interne revient alors au même. Le deuxième cas correspond à une fonctionnement différentiel de *certains* items selon les groupes. Tester l'adéquation revient alors à comparer la courbe caractéristique des items dans chaque groupe, autrement dit à tester une hypothèse de biais. Supposons que le test soit négatif. Rien ne prouve qu'il le serait encore si l'on divisait l'échantillon selon un autre critère (*cf.* à ce sujet van den Wollenberg, 1982, b, p. 138). Le modèle de Rasch n'a pas d'autre prétention que de permettre l'obtention de paramètres d'item indépendants *de la distribution des mesures des sujets dans le trait latent.* La question de savoir si le modèle tient dans des groupes de sujets différents est une question empirique qui ne peut recevoir de réponse générale sur la base d'une comparaison unique (cf. Goldstein et Wood, 1989, p. 155).

En résumé, on se trouve devant un problème sans solution : le score total n'est pas un bon critère, mais on ne connaît pas de solution de rechange efficace.

2.6. Les contraintes exercées par les logiciels

Les tests sont nombreux (Hambleton, 1989, p. 175, évoque avec une certaine exagération "a myriad of statistical tests"), mais ils sont tous imparfaits et sont inégalement efficaces. Le chercheur est donc contraint à la fois de faire un choix et d'utiliser plusieurs tests. Mais son choix n'est pas libre. Il dépend très étroitement du logiciel employé. Du point de vue des tests d'adéquation, les programmes sont inégalement performants. NORHAM (Fraser & McDonald, 1988), par exemple, comporte trois tests d'adéquation dont un seul paraît satisfaisant d'après l'étude de Rogers et Hattie (1967). PML-PC (Molenaar, 1990) offre sept tests différents, dont quatre excellents. En revanche, il ne permet pas comparer plusieurs modèles, comme on peut le faire avec BILOG (Mislevy & Bock, 1990). Ces limites conduisent à préconiser l'utilisation conjointe d'au moins deux logiciels complémentaires, estimant les paramètres par des méthodes différentes et offrant au total un éventail suffisamment large de tests d'adéquation. Le couplage de BILOG et de PML-PC nous paraît particulièrement heureux. Les tests d'adéquation proposés par les logiciels peuvent ainsi être complétés par certaines méthodes graphiques d'évaluation de l'adéquation (*cf.* partie 3).

2.7. L'incidence de l'estimation des paramètres sur la mesure de l'adéquation.

L'inadéquation des données aux modèles a généralement pour cause le non respect par l'échelle de mesure d'une ou plusieurs des conditions d'application du modèle : l'échelle mesure de deux ou plusieurs concepts différents, les items n'ont pas la même discrimination, les réponses aux items ne sont pas indépendantes les unes des autres, etc.. Mais l'inadéquation peut provenir également d'une mauvaise estimation des paramètres (McKinley & Mills, 1985). Un exemple de l'effet de l'estimation des paramètres sur la mesure de l'adéquation est donné plus loin (cf. § 3.3, test de Lord). On doit donc tenir compte dans l'évaluation de l'adéquation de la qualité de l'estimation des paramètres, qui dépend de nombreux facteurs : la méthode d'estimation

choisie, les modifications de l'algorithme originel dues aux contraintes informatiques[5], le nombre d'items, le nombre de sujets, la distribution de la compétence des sujets (Stocking, 1990)... Le chercheur ne peut donc s'en remettre entièrement aux tests d'adéquation. Une familiarité avec le modèle et le logiciel sont requises.

3. LES TESTS D'ADÉQUATION

Cette revue est limitée aux tests portant sur les items, car la question que l'on se pose en général est de savoir si les items forment une échelle raschienne (*item fit*) et rarement celle de savoir s'il les courbes caractéristiques des sujets[6] sont conformes au modèle (*person fit*) ; beaucoup de tests peuvent d'ailleurs être employés dans les deux perspectives (par sommation des résidus sur les personnes ou sur les items selon le cas). «Test d'adéquation» est pris dans un sens large : l'expression désigne non seulement les tests statistiques, mais également les méthodes graphiques.

Les tests d'adéquation peuvent être répartis en trois grandes classes correspondant à des prédictions différentes déduites du modèle. L'écart entre les données et les prévisions mesure le degré d'inadéquation au modèle. La première classe regroupe les tests statistiques ou graphiques qui comparent la fréquence des réponses exactes observée et la fréquence attendue d'après le modèle. Cette classe peut être subdivisée en deux sous-classes selon que les différences résiduelles sont calculées au niveau de chaque item considéré isolément ou au niveau de chaque paire d'items. Traub et Lam (1985) appellent les premiers "indicateurs de premier ordre" et les autres "indicateurs de deuxième ordre". Comme nous le verrons plus loin, les tests de deuxième ordre permettent d'échapper à une limitation fondamentale des tests de premier ordre. La deuxième classe est constituée par les tests de rapport de vraisemblance (tests statistiques uniquement). Un test classique de cette catégorie est celui d'Andersen (1973, b) qui décompose le maximum de la fonction de vraisemblance en un produit de maximums, calculés sur des sous-échantillons disjoints, et le compare au maximum de vraisemblance total, auquel il doit être égal selon le modèle. La troisième classe est celle des tests (statistiques ou graphiques) où l'on subdivise l'échantillon en deux et où l'on compare la valeur des paramètres dans les deux sous-échantillons. En raison du principe d'invariance des paramètres, la localisation des items doit être la même dans les deux sous-échantillons si les données sont conformes au modèle.

Notre recension s'étend à vingt-quatre tests d'adéquation. Bien qu'elle soit très extensive, elle ne prétend pas être exhaustive. Certains tests ont pu échapper à nos investigations, tandis que quelques tests non publiés ont délibérément été omis[7]. Pour chaque test, on indique son fondement théorique, s'il n'est pas apparent, ainsi que sa formule s'il s'agit d'un test statistique, et l'on fournit si possible des éléments d'évaluation. Les observations générales faites dans la deuxième partie, qu'il convient de ne pas perdre de vue, ne sont qu'exceptionnellement reprises. A l'occasion, on indique où trouver un exemple d'application détaillé.

Pour faciliter la lecture des formules, nous nous sommes efforcé de simplifier et d'homogénéiser la notation. S'il se reporte aux publications originales, le lecteur ne devra donc pas être surpris de constater des différences d'écriture. Dans le texte et les formules, *les indices s et i désignent respectivement le sujet et l'item ; k est le nombre d'items, r est le score total du*

---

[5] Baker (1987) attire l'attention sur ce point et signale, à titre d'exemple, que BICAL divise par deux la localisation de l'item quand sa valeur devient excessive.

[6] La courbe caractéristique d'un sujet localisé en θ^* est la courbe représentative de P_{si} pour $\theta = \theta^*$ qui est une fonction décroissante b (la localisation de l'item).

[7] Il s'agit de deux tests pour lesquels nous ne disposons pas d'informations suffisantes (Mead, 1976, cité par Gustafsson, 1980, et Hambleton & Swaminathan, 1985; Stene, 1969, cité par Molenaar, 1963) et d'un test de validité fort douteuse (test de Elliott, Murray & Saunders, 1977, cité et analysé par Yen, 1981).

sujet (ou le sous-échantillon formé des sujets ayant le score total r) et N le nombre total de sujets.

3.1. Tests fondés sur les différences résiduelles

3.1.1 Tests de premier ordre

A) Tests graphiques

1°. Méthode de Rasch (1960/1980)

Appelons θ_r la localisation des sujets ayant un score total r, P_{ri} la probabilité de réussite à l'item i pour un sujet localisé en θ_r et Q_{ri} la probabilité complémentaire de P_{ri}. D'après le modèle de Rasch (cf. formule 2), *log* (P_{ri} / Q_{ri}) est, quand b_i est fixé, une fonction linéaire de θ_r *log* $(P_{ri} / Q_{ri}) = \theta_r - b_i$. La droite représentant cette fonction coupe l'axe des abscisses au point $(b_i, 0)$ et l'axe des ordonnées au point $(0, -b_i)$. On peut donc facilement la tracer à partir des estimateurs de θ_r et de b_i. Les points d'intersection $(b_i, 0)$ et $(0, -b_i)$ montrent que toutes les droites théoriques sont parallèles. Ce parallélisme correspond à la constance de la discrimination des items dans le modèle de Rasch. Les droites théoriques ne sont d'ailleurs que la linéarisation des courbes caractéristiques d'item.

Le test d'adéquation proposé par Rasch (1960/1980) consiste à étudier comment se répartissent, par rapport aux droites théoriques, les points $(\theta_r, log\ (p_{ri}/q_{ri}))$, p_{ri} étant la proportion observée de sujets de score r qui donnent la réponse codée 1 à l'item i. Le graphique permet de juger les écarts entre la fréquence des réponses 1 à l'item et la fréquence attendue de ces réponses pour chaque score r. Il permet également d'évaluer dans quelle mesure la condition de discrimination constante est satisfaite par les items.

Cette méthode, simple d'emploi et riche en information, mériterait d'être davantage employée.

2°. Mise en relation graphique des scores observés et prédits

Divers auteurs (par ex. Gustafsson, 1980, b ; Whitely, 1980) proposent une méthode graphique simple dans son principe. Pour chaque item i, on porte en abscisse la proportion p_{ri} de réponses codées 1 observées dans chaque groupe de score total r, et en ordonnée, la proportion P_{ri} correspondante prédite à partir du modèle. Pour calculer P_{ri}, on utilise l'estimation de la valeur de θ correspondant à r. Si l'item est conforme au modèle, les points sont alignés sur une droite passant par l'origine. Si l'item a une discrimination trop faible par rapport aux autres, on observe la relation $p_{ri} > P_{ri}$ pour les valeurs faibles de r, et $p_{ri} < P_{ri}$ pour les valeurs élevées de r. Si l'item a une discrimination trop forte, le pattern inverse s'observe ($p_{ri} < P_{ri}$ pour les valeurs faibles de r et $p_{ri} > P_{ri}$ pour les valeurs élevées).

Les auteurs reconnaissent que cette méthode souffre d'une certaine subjectivité, mais elle permet de comprendre certaines causes d'inadéquation.

3°. Méthode de Hambleton

Hambleton (Hambleton & Swaminathan, 1985 ; Hambleton, Swaminathan & Rogers, 1991) propose de construire pour chaque item un graphique où les résidus standardisés[8] sont étudiés en fonction de la localisation des sujets. Ceux-ci sont regroupés en classes selon leur mesure dans la dimension latente et la moyenne des résidus de chaque classe est portée en

---

[8] Les résidus standardisés z_{ci} sont les résidus bruts divisés par l'erreur standard de la probabilité prédite:

$$z_{ci} = (p_{ci} - P_{ci}) / (P_{ci}\ Q_{ci}\ /N_c)^{1/2}$$

où c désigne la classe.

ordonnée. L'auteur préconise de 10 à 15 classes, afin que l'effectif de chacune soit suffisant et que l'intervalle de variation de θ soit assez faible pour que la classe puisse être tenue pour homogène.

Ces graphiques offrent deux intérêts. Le premier est de permettre de faire des hypothèses sur les causes d'inadéquation. Une discrimination trop forte ou trop faible par exemple est identifiable par une forte dispersion des résidus et des signes opposés pour les valeurs de θ faibles ou élevées. Le deuxième intérêt est de permettre de comparer l'adéquation à différents modèles. Si, par exemple, le modèle à deux paramètres est nettement plus adapté, les résidus sont plus faibles, moins dispersés et ne font pas apparaître de pattern identifiable. L'auteur fournit des exemples tout à fait convaincants. On remarquera cependant que la méthode présente d'autant moins d'intérêt que le nombre d'items est faible.

4°. Méthode de Ludlow (1985, 1986)

Dans la méthode proposée par Ludlow (1985, 1986), la distribution spatiale des résidus est comparée à celle de résidus aléatoires obtenus à partir d'une simulation. Cette confrontation des résidus réels et des résidus aléatoires permet d'éviter les erreurs de jugement dues aux fluctuations aléatoires, qui peuvent être importantes. Les données sont générées à partir des estimateurs des paramètres d'item et de compétence («*tailored simulation*»). Les résidus aléatoires sont ainsi calculés très finement. Pour éviter les aléas de la génération de données, plusieurs simulations ont lieu et on ne retient que les déviations retrouvées d'une simulation à l'autre.

La méthode de Ludlow se caractérise également par une stratégie qui multiplie et hiérarchise les analyses, et qui permet de diagnostiquer les causes d'inadéquation[9]. Après une comparaison globale, où l'on compare le nombre de résidus supérieurs à une valeur prise comme critère dans les données réelles et les données simulées, on construit plusieurs graphiques, où les résidus standard sont mis en relation successivement avec la localisation des individus, la difficulté et l'ordre des items. Le but recherché est d'identifier la ou les raisons des inadéquations constatées. Par exemple, un graphique où les items sont ordonnés comme ils le sont dans l'échelle peut révéler une violation de la condition d'indépendance locale due à un phénomène d'apprentissage. Il n'est pas possible ici d'exposer plus avant cette procédure, qui est décrite en détail dans l'article de 1985 et exemplifiée dans celui 1986.

La méthode de Ludlow est très intéressante, mais son application est limitée par l'absence de programme.

B) Tests statistiques

B-1 Tests du Khi-deux

Plusieurs tests très employés sont construits de la même manière. Les sujets sont regroupés en C classes en fonction de leur mesure estimée dans la dimension latente. Pour un item i, les réponses sont disposées dans un tableau (2 × C). Les fréquences observées dans les cellules du tableau sont comparées aux fréquences théoriques résultant du modèle au moyen du test du Khi-deux. Un test global est obtenu en sommant les valeurs obtenues pour les k items. Les tests diffèrent les uns des autres par le mode de construction des classes (fractiles ou intervalles égaux), le nombre de classes et le calcul des fréquences théoriques. En raison de sa fréquence d'emploi, nous commencerons par le test de Yen (1981), que nous exposerons plus en détail, les autres pouvant être présentés ensuite plus brièvement.

---

[9] Hambleton, Swaminathan et Rogers, 1991 présentent eux aussi une méthode graphique où les résidus observés sont comparés aux résidus aléatoires issus d'une simulation. Mais ces auteurs se contentent de comparer les histogrammes des kN résidus standards.

5°. Test Q_1 de Yen (1981)

Dans le test de Yen (1981), on ordonne les sujets selon leur localisation estimée, puis on les répartit en 10 classes d'effectifs sensiblement égaux. Considérons un item i. La probabilité théorique pour qu'un sujet fournisse la réponse observée u_{si} est donnée par le modèle et vaut P_{si} si $u_{si} = 1$ et Q_{si} si $u_{si} = 0$. Pour l'obtenir, on remplace θ par son estimateur. La fréquence théorique d'une réponse pour une classe c de sujets est, dans ce test, la moyenne des n_c probabilités individuelles. Le test Q_{1i} compare les fréquences observées et les fréquences théoriques dans les 20 cellules de la matrice des réponses à l'item :

$$Q_{1i} = \Sigma_{10}\ n_c\ (p_{ci} - P_{ci})^2 / P_{ci}\ + \Sigma_{10}\ n_c\ (q_{ci} - Q_{ci})^2 / Q_{ci}\ , \qquad [12]$$

où p_{ci} et q_{ci} désignent les proportions observées de réponses respectivement codées 1 et 0. La formule [12] peut s'écrire plus simplement :

$$Q_{1i} = \Sigma_{10}\ n_c\ (p_{ci} - P_{ci})^2 / P_{ci}\ Q_{ci}. \qquad [13]$$

Q_{1i} est distribué, selon l'auteur, approximativement comme un Khi-deux à (10-1) degrés de liberté. Yen discute en détail la question du nombre de degrés de liberté. Les degrés de liberté perdus reflètent la dépendance du calcul des fréquences théoriques par rapport aux données. L'auteur considère que l'estimation de la localisation de l'item doit entraîner la perte d'un degré de liberté, mais qu'il en va différemment pour la localisation des sujets. En effet celle-ci est fondée sur les réponses à l'ensemble des items, de sorte que, si l'échelle est suffisamment longue, la contribution de chaque item à l'estimation de θ est faible. Yen fait observer que la détermination des limites de classe est dépendante de la distribution de la localisation estimée des sujets, ce qui doit affecter le nombre de degrés de liberté, mais elle ne précise pas davantage cet effet.

Cette discussion met en évidence plusieurs points. En premier lieu, elle illustre bien l'incertitude qui existe sur la distribution des statistiques d'adéquation de ce type. En deuxième lieu elle montre que l'approximation de la distribution repose parfois sur des hypothèses non immédiatement évidentes (ici, celle relative au nombre d'items). En troisième lieu elle éclaire l'origine des débats entre les auteurs et fait ressortir la nécessité de contrôler les suppositions théoriques par des études de simulation. Ces simulations sont d'autant plus nécessaires que l'étude théorique de la statistique omet parfois certains points ; Yen par exemple n'examine pas la question de l'influence possible du groupement en classes sur la mesure de l'adéquation.

McKinley & Mills (1985) effectuent une étude de simulation. Selon les conditions, les données générées sont conformes au modèle de Rasch, au modèle logistique à deux paramètres, au modèle logistique à trois paramètres ou à un modèle bi-dimensionnel. Les auteurs font varier par ailleurs la taille de l'échantillon (500, 1000 ou 2000 sujets) et la compétence moyenne des sujets (trois niveaux distingués), soit neuf échantillons. Le test de Yen apparaît comme fiable pour un effectif de 1000 sujets, aussi bien en ce qui concerne les erreurs de type I (rejet du modèle alors que les données y sont conformes) que les erreurs de type II (acceptation du modèle alors que les données s'en écartent). Les résultats de McKinley et Mills sont en accord avec ceux obtenus par Yen elle-même dans une simulation analogue ; on notera que le nombre d'items est très différent dans les deux études (respectivement 75 et 36).

Un des intérêts du test de Yen est qu'il permet de comparer l'adéquation des données au modèle de Rasch et au modèle à deux paramètres. Il suffit selon l'auteur de comparer la valeur moyenne de Q_{1i} pour l'un et l'autre modèle. Le modèle de Rasch convient si ces moyennes sont «approximativement égales» (p. 261), mais l'auteur ne précise pas davantage la limite acceptable de cette différence.

6°. Test de Bock (1972)

Le test de Bock (1972) est proche du test de Yen. Il en diffère par le nombre C de classes, qui n'est pas fixé à l'avance, et surtout par le calcul des fréquences théoriques P_{ci} et Q_{ci} dans les diverses classes c. Ces fréquences s'obtiennent en donnant comme valeur à θ la médiane de l'intervalle de variation de la classe c. La formule de l'indice de Bock est donc :

$$\chi_i^2 = \Sigma_c \; n_c \, (p_{ci} - P_{ci})^2 / \, P_{ci} \, Q_{ci} \qquad [14]$$

avec les mêmes symboles que précédemment (n_c est le nombre de sujets de la classe c). Le nombre de degrés de liberté est (C-1). Les simulations faites par McKinley & Mills (1985) font apparaître peu de différence entre le test de Bock et le test de Yen.

7°. Test Y de Wright-Panchapakesan (1969)

Le test de Wright et Panchapakesan (1969) se caractérise par le groupement des sujets en classes, qui s'effectue sur la base du score total r. Les classes ont un intervalle constant ; en outre, si l'échelle est moyenne ou longue, elles sont plus nombreuses que dans les deux tests précédents. Les fréquences théoriques P_{ri} et Q_{ri} se calculent plus facilement que dans les deux tests précédents puisque r étant une statistique suffisante pour θ, les sujets de même score ont la même mesure dans le trait latent. Le test a pour formule :

$$\chi_i^2 = \Sigma_r \; n_r \, (p_{ri} - P_{ri})^2 / \, P_{ri} \, Q_{ri} \,, \qquad [15]$$

avec $0 < r < k$, puisqu'on ne sait pas estimer par le maximum de vraisemblance la localisation des sujets de score nul ou parfait. Selon les auteurs le nombre degré de liberté est égal à (k-2) si tous les scores sont représentés. Par sommation des χ_i^2, on obtient un test global :

$$Y = \Sigma_i \; \chi_i^2 \qquad [16]$$

distribué asymptotiquement comme un χ^2 à (k-1)(k-2) degrés de liberté.

Ce test a été critiqué par plusieurs auteurs, notamment Divgi (1981) et van den Wollenberg (1982, b), sur la base d'analyses aussi bien théoriques qu'empiriques. On lui reproche en particulier d'occasionner des erreurs de type I quand k est petit. Mais le test a aussi des avocats. Yen (1981) souligne la proximité de ce test et du sien. Andrich (1988) montre sur une échelle courte (six items) qu'il donne des indications proches du test d'Andersen[10].

8°. Khi-deux de Wright, Mead & Bell (1979)

Wright et Mead (cf. Wright, Mead & Bell, 1979) proposent un autre test où les sujets sont regroupés sur la base de leur score total r en C classes, de façon à ce que la distribution soit approximativement uniforme et que C soit égal ou inférieur à 6. La formule en est :

$$V_{Bi} = 1/(C-1) \, [\Sigma_c \; (x_{ci} - \Sigma_r \, n_r \, P_{ri})^2 \, / \, (\Sigma \; n_r \, P_{ri} \, Q_{ri})] \; (k/k-1) \qquad [17]$$

où x_{ci} est le nombre de sujets de la classe c donnant la réponse 1, n_r est l'effectif des sujets c obtenant le score total r, P_{ri} est la probabilité de répondre 1 à i pour un sujet de score r. Le nombre de degrés de liberté est égal à (C-1).

---

[10] L'argument est affaibli par le fait que l'échelle n'est pas conforme au modèle et que le test de Wright & Panchapakesan apparaît, d'après cet exemple, plus puissant que le test d'Andersen, lui-même jugé trop puissant quand N est élevé (ici, N= 1482).

Les simulations de McKinley et Mills (1985) montrent que ce test donne des résultats proches de ceux de Yen et de Bock ; celles Rogers et Hattie (1987) indiquent que le test est sensible aux différences de discrimination mais beaucoup moins à la multidimensionnalité des données.

B-2 Tests fondés sur les résidus standards

Wright et ses collaborateurs ont proposé deux autres tests d'adéquation fondés sur les résidus standards.

9°. Test de Wright & Stone (1979)

A la différence des tests précédents où les sujets sont groupés en classe, le test de Wright & Stone (1979) considère la réponse de chaque sujet à chaque item. L'espérance mathématique d'une réponse u_{si} ($u_{si} = 1$ ou $u_{si} = 0$) est la probabilité P_{si} fournie par le modèle, que l'on peut estimer en remplaçant θ_s et b_i par leurs estimateurs ; la variance de u_{si} est égale à $P_{si}\, Q_{si}$. Les résidus standards $(u_{si} - P_{si}) / (P_{si}\, Q_{si})^{1/2}$ suivent approximativement la loi normale réduite selon les auteurs. En élevant ces résidus z_{si} au carré et en sommant sur les N individus, on obtient un test d'adéquation de l'item, distribué selon la loi du Khi-deux à (N-1) degrés de liberté. Le résidu carré moyen résulte d'une division par le nombre de degré de liberté :

$$F = (\Sigma_s \; z_{si}^2) / (N-1), \qquad [18]$$

distribué comme un F avec $(N-1, \infty)$ degrés de liberté.

Rogers et Hattie (1987) contestent vigoureusement l'approximation de z_{si} par la loi normale. Les simulations faites par ces auteurs montrent que le test conduit à des erreurs de type I et des erreurs de type II trop fréquentes.

10°. Test "t total" de Wright, Mead & Bell (1979)

Ce test est une variante du précédent. La somme des carrés des résidus est pondérée par la somme des variances des résidus :

$$v_i = (\Sigma_s \; z_{si}^2) / (\Sigma_s \; P_{si}\, Q_{si}) \qquad [19].$$

Cette statistique est appelée "test t total" par Wright, Mead et Bell (1979). V aurait pour espérance 1 et pour variance $(\Sigma w - 4\Sigma w^2)/(\Sigma w)^2$, avec $w = P_{si}\, Q_{si}$ (Wright, 1985). Sa distribution n'est pas connue, mais les auteurs conjecturent que la racine cubique de v a une distribution asymptotique normale.

Les simulations de Rogers et Hattie (1987) font apparaître peu d'erreurs de type I, mais beaucoup trop d'erreurs de type II.

B-3 Tests de Martin-Löf et de van den Wollenberg

Ces deux tests ont comme particularité de s'appliquer au cas où les paramètres sont estimés par le maximum de vraisemblance conditionnel (CML).

11°. Test T de Martin-Löf (1973)

Martin-Löf ayant présenté son test dans un rapport non publié (1973) rédigé en suédois, nous référons aux articles de Gustafsson (1980), Molenaar (1983) et van den Wollenberg (1982, b,).

Martin-Löf considère les (k-1) groupes de n_r sujets ayant le même score r. La fréquence attendue des réponses codées 1 dans un de ces groupes est $n_r\ P_{ri}$, avec

$$P_{ri} = b_i \gamma_r' / \gamma_r\ , \qquad [20]$$

où b_i est la difficulté estimée de l'item, γ_r est la fonction symétrique élémentaire d'ordre r (cf. § 1.5) et γ_r'sa dérivée.

Le test proposé est

$$T = \Sigma_r\ \mathbf{d}'_r\ \mathbf{V}_r^{-1}\ \mathbf{d}_r \qquad [21]$$

où $\mathbf{d}_r$ est le vecteur des k différences entre les fréquences observées et attendues des réponses codées 1 et où $\mathbf{V}_r$ est la matrice (k x k) de variances-covariances des fréquences attendues. T est distribué comme un χ^2 à (k-1)(k-2) degrés de liberté quand chaque n_r tend vers l'infini (Gustafsson déconseille l'utilisation de T si $n_r < 10$ dans un groupe).

Bien que l'on puisse analyser la contribution au Khi-deux de chaque groupe r, T est avant tout un test global, qui, comme tous les tests de cette catégorie pose un problème de puissance. Selon Molenaar (1990, p.39) «a highly significant outcome need not indicate a serious misfit». D'après les simulations de van den Wollenberg (1982, b) le test de Martin-Löf donne des résultats proche de celui d'Andersen (cf. *infra*).

12°. Test Q_1 de van den Wollenberg (1982,b)

Le test Q_1 de van den Wollenberg (1982, b) compare, dans chaque item et chaque groupe r, la fréquence observée et attendue des réponses. Le calcul de $n_r\ P_{ri}$, fréquence attendue des réponses codées 1, est exactement le même que dans le test T de Martin-Löf. Pour un item, le Khi-deux vaut

$$q_i = \Sigma_r\ [(n_{ri} - n_r\ P_{ri})^2\ /\ n_r P_{ri}] + [(n_{ri} - n_r P_{ri})^2\ /\ (n_r - n_r\ P_{ri})] \qquad [22].$$

Q_1 s'obtient par sommation des q_i ; un facteur de correction est introduit pour tenir compte de la perte d'un degré de liberté :

$$Q_1 = (k\text{-}1)/k\ \ \Sigma_i\ q_i \qquad [23],$$

distribué comme un Khi-deux à (k-1)(k-2) degrés de liberté.

Les simulations de l'auteur montre que Q_1 est très proche du T de Martin-Löf (Q_1 est d'ailleurs exactement égal à T dans un cas particulier) ainsi que du test d'Andersen. Il présente donc les mêmes avantages et les mêmes défauts que ces deux tests classiques. On peut dès lors s'interroger sur l'utilité de ce test supplémentaire. Les raisons d'ordre pratique fournies par l'auteur, ne sont pas vraiment convaincantes. Ce test, qu'il ne faut pas confondre avec le Q_1 de Yen, est peu employé.

B-4 Tests de Molenaar

13°. Test binomial (Molenaar, 1983)

Le test binomial de Molenaar (1983) est un test ultra-analytique qui utilise une distribution exacte pour prévenir les problèmes posés par l'inévitable faiblesse des effectifs de certains groupes.

Pour chaque item i et chaque groupe r, on calcule la probabilité P_{ri} des réponses codées 1, de la même façon que dans le test T de Martin-Löf. Si les données sont conformes au modèle, la fréquence attendue $n_r P_{ri}$ suit la loi binomiale. On peut ainsi comparer par le test binomial exact la fréquence réelle des réussites, n_{ri}, à la fréquence théorique $n_r P_{ri}$.

Le nombre des comparaisons est égal à k(k-1). Il devient rapidement excessif et gêne l'interprétation. Mais il est possible de repérer des structures signifiantes. Molenaar observe que si un item est faiblement corrélé aux autres (faible discrimination), la probabilité d'une réponse correcte augmente peu avec r et que l'on peut s'attendre à $n_{ri} > n_r P_{ri}$ pour les faibles valeurs de r et à $n_{ri} < n_r P_{ri}$ pour les fortes valeurs. Si l'item est fortement corrélé aux autres (forte discrimination), les relations inverses sont prévisibles. On a donc le moyen de repérer une cause fréquente d'inadéquation. Des exemples sont donnés par l'auteur.

Un autre inconvénient du test est la fréquence des erreurs de première espèce dans les échelles longues (Molenaar, 1990). C'est pour y remédier que l'auteur propose le test qui suit.

14°. Test U de Molenaar (1983)

Le test U de Molenaar vise à prévenir le risque de capitalisation des chances encouru avec le test binomial. D'un test ultra-analytique, on passe à un test analytique.

Le test compare les résidus pour les sujets ayant un score total r faible ($r \in L = \{1, 2,..., r_1\}$) aux résidus pour les sujets ayant un score total élevé ($r \in R = \{r_2, r_2+1, ..., k-1\}$). Le test U s'écrit :

$$U = (\Sigma_L \ Z_{ri} - \Sigma_R \ Z_{ri}) / (r_1 + k - r_2)^{1/2}, \qquad [24]$$

où Z_{ri} est le résidu standard du groupe r :

$$Z_{ri} = (n_{ri} - n_r P_{ri}) / (n_r P_{ri} Q_{ri})^{1/2}) \qquad [25].$$

Selon l'auteur, U se distribue suit une loi quasi normale (quand, $\forall r$, $n_r \rightarrow \infty$). Un bon usage de U suppose $n_r > 30$ pour toutes les valeurs de r ; cette condition s'impose plus encore pour les groupes où P_{ri} est voisin de 0 ou de 1.

Une valeur de U positive suggère que l'item a une discrimination trop faible, une valeur négative, qu'il a une discrimination plus élevée que la moyenne.

3.1.2 Tests de deuxième ordre

Les tests de deuxième ordre visent à pallier l'insuffisante sensibilité des tests classiques à la violation de l'hypothèse d'unidimensionnalité (cf. § 2.1). Les deux tests présentés supposent les paramètres estimés par le maximum de vraisemblance conditionnel. Leur intérêt est surtout théorique.

15°. Test Q_2 de van den Wollenberg (1982, b)

Si les données sont unidimensionnelles, la corrélation entre deux items s'annule quand les sujets ont la même mesure dans la dimension latente et donc, dans le modèle de Rasch, le même score total r. Le principe du test de van den Wollenberg (1982, b) consiste à vérifier cette indépendance des réponses des sujets de même compétence.

Pour chaque paire d'items i et j et chaque groupe de score r, on dresse le tableau de contingence ; on note n_r l'effectif du groupe et n_{rij} celui des sujets qui répondent 1 à la fois à i et

à j. Le calcul de Q_2 comporte trois phases. La première étape est le calcul de l'espérance de $(n_{rij} \mid n_r)$:

$$E(n_{rij} \mid n_r) = n_r \, (b_i \, b_j \, \gamma_{r-2}'') \, / \, \gamma_r \, , \qquad [26]$$

où b_i et b_j désignent la localisation des items estimée dans le groupe r, γ_r est la fonction symétrique élémentaire d'ordre r et à γ_{r-2} "la dérivée seconde de cette fonction. Les fréquences attendues dans les trois autres cases du tableau s'obtiennent par différence. La deuxième étape est le calcul d'un Khi-deux ordinaire, noté Q_{rij}. La troisième étape est une sommation des $k(k-1)/2$ Q_{rij}, avec un facteur de correction tenant compte de la non indépendance paires d'items (et donc des termes de la somme) :

$$Q_{2(r)} = (k-3)/(k-1) \; \Sigma_i \, \Sigma_j \; Q_{rij} \qquad [27].$$

$Q_{2(r)}$ se distribue comme un Khi-deux à $k(k-3)/2$ degrés de liberté.

L'intérêt de ce test est de permettre un diagnostic d'unidimensionnalité dans les cas où les autres tests échouent. Les simulations de l'auteur prouvent la supériorité de Q_2 sur des tests aussi puissants que les tests de Martin-Löf et d'Andersen. Mais ce n'est pas un test d'adéquation complet. Cette limite doit être mise en rapport avec la complexité des calculs requis : la localisation des items doit être estimée dans chacun des (k-1) groupes de sujet, ceci afin d'assurer l'égalité des fréquences *marginales*, théoriques et observées. Le test Q_1 a donc une valeur essentiellement théorique[11].

16°. Test d'unidimensionnalité de Molenaar (1983)

Le principe de base du test de Molenaar (1983) est le même que celui du test Q_2. Mais Molenaar considère que le calcul de l'espérance de n_{rij} doit se faire sous la triple condition de n_{ri}, n_{rj}, et n_r, et pas seulement sous celle de n_r.

L'auteur considère la question de savoir si n_{rij} dévie significativement de son espérance conditionnelle, les totaux marginaux étant fixés. Il établit que $P(n_{rij} \mid n_{ri}, n_{rj},n_r)$ suit la loi hypergéométrique étendue. Cette loi est utilisée dans la comparaison des fréquences observées et attendues. Nous renvoyons à l'article très technique de Molenaar pour des précisions.

Les remarques faites au sujet du test de van den Wollenberg s'appliquent à celui de Molenaar. Ce test n'est pas disponible sur PML-PC, programme auquel l'auteur a contribué et dont la version actuelle (1990) est pourtant postérieure de sept ans à l'article. Peut-être est-ce un signe du faible intérêt pratique du test au regard de sa complexité

3.2. Tests du rapport de vraisemblance

Les tests du rapport de vraisemblance consistent à calculer le rapport λ entre le maximum de la fonction de vraisemblance sous une hypothèse et le maximum de cette fonction sous l'hypothèse alternative (Waller, 1981). Si le nombre d'observations est grand, $-2 \, log \, \lambda$ se distribue comme un Khi-deux. Whitely (1980) estime que les tests du rapport de vraisemblance sont les plus sûrs, car ils sont dérivés de la procédure d'estimation des paramètres.

---

11 Une discussion théorique du test de van den Wollenberg dépasse le cadre de cet article. On peut se reporter aux remarques de l'auteur (p. 139), aux critiques de Molenaar (1983, p. 66-67) et à la réponse de van den Wollenberg (p. 136-137). Un problème non évoqué par ces deux auteurs est celui de la combinaison des résultats des divers $Q_{2(r)}$.

17°. Test d'Andersen (1973, b)

Andersen (1973, b) teste l'adéquation en rapportant le maximum de vraisemblance pour l'échantillon total (V_t) au produit des maximums de vraisemblance de g groupes disjoints (ΠV_g). Andersen montre que

$$Z = -2 (log\ V_t - \Sigma_g\ log\ V_g) \qquad [28]$$

se distribue asympotiquement comme un Khi-deux avec (k-1) (g-1) degrés de liberté, quand ($\forall$ g) $n_g \rightarrow \infty$. Si les groupes correspondent aux divers scores totaux r, Z possède (k-1) (k-2) degrés de liberté. On dichotomise très souvent l'échantillon selon r ; Z a alors (k-1) degrés de liberté.

Les auteurs s'accordent à reconnaître que le test d'Andersen est l'un des mieux fondé, mais il passe pour être trop puissant quand le nombre des sujets N est grand. Pour Molenaar (1990, p.38) «P-values between .001 and .05 may not indicate serious misfit». Mais la puissance du test ne dépend pas que de N ; elle est aussi fonction du nombre d'items et de la variance de la mesure des sujets dans la dimension latente. Pour un effectif de quelques centaines de sujets le test peut, dans certaines conditions, ne pas être assez puissant (Gustafsson 1980, b ; Whitely, 1980).

18°. Test ML-PCC de Martin-Löf (1973)

Martin-Löf (1973) a conçu un deuxième test, moins connu que le test T présenté plus haut, mais qui permet des comparaisons impossibles avec les autres tests.

Les k items sont répartis en deux sous-ensembles disjoints k_A et k_B. Les items sont calibrés par le CML d'abord globalement (maximum de vraisemblance : V_t) puis séparément pour chacun des deux sous-ensembles k_A (maximum : V_A) et k_B (maximum : V_B). On calcule la quantité

$$log\lambda = -\Sigma r_A\ \Sigma r_B\ n_{rArB}\ log(n_{rA}\ n_{rB})/n + \Sigma r\ n_r\ log\ n_r/n + (V_t - V_A - V_B), \qquad [29]$$

où r_A et r_B désignent les scores totaux dans la première et deuxième sous-échelle respectivement et où n_{rArB} indique le nombre de sujets ayant un score total r_A dans A et r_B dans B. Martin-Löf montre que ML-PCC= -2 log λ se distribue comme un χ avec ($k_A\ k_B$ - 1) degrés de liberté.

Divers modes de subdivision permettent de tester diverses hypothèses. En groupant les items selon leur difficulté, on teste l'hypothèse de constance de la discrimination (Gustafsson, 1980). On peut aussi comparer des items situés en début et fin d'échelle pour tester un effet d'apprentissage ou l'existence d'un facteur de vitesse étranger à la dimension théoriquement mesurée par l'échelle. Le test est particulièrement utile quand les items peuvent être subdivisés selon un critère tel que la forme ou le contenu des items. Il permet alors de vérifier que ces sous-échelles mesurent la même dimension latente[12].

19°. Test G^2 de Mislevy & Bock

G^2 est un test d'hypothèse classique (cf. Bishop, Fienberg & Holland, 1975) qui peut être largement utilisé (modèles diférents, méthodes d'estimation différentes). Mislevy et Bock l'ont introduit dans le programme BILOG (1982) qui estime les paramètres par le maximum de vraisemblance marginal. Nous utilisons pour présenter ce test le manuel de la version 3 de BILOG (Mislevy & Bock, 1990).

---

[12] Il ne s'agit pas toutefois d'un véritable test d'unidimensionnalité, car la différence éventuellement révélée peut provenir d'une différence de discrimination moyenne des items de A et de B, alors même qu'ils mesurent une dimension latente unique.

La formule de G^2 diffère selon le nombre d'items.

Si $k \leq 10$ et si les données comprennent la plupart des 2^k vecteurs de réponse possibles, G^2 fournit un test global qui s'écrit :

$$G^2 = 2 \sum f_l \, log \, (r_l / N \, P(\mathbf{u}_l)), \qquad [30]$$

où f_l est la fréquence du vecteur $\mathbf{u}_l$ ($l \in \{1,2^k\}$) et $P(\mathbf{u}_l)$ la probabilité marginale de $\mathbf{u}_l$. G^2 est distribué comme un Khi-deux à $(2^k - k - 1)$ degrés de liberté.

Si $k > 20$, il est possible de regrouper les sujets en classes en fonction de leur mesure dans la dimension latente. En désignant chaque intervalle de θ par c, le test s'écrit :

$$G_i^2 = 2\sum_c [n_{ci} \, log \, (n_{ci} / n_c \, P_{ci}) + (n_c - n_{ci}) \, log \, (n_c - n_{ci} / n_c \, Q_{ci})], \qquad [31]$$

où n_{ci} est le nombre de réponses codées 1 à l'item i dans la classe c, n_c est l'effectif de c et P_{ci} la probabilité théorique d'une réponse 1 à i pour une valeur de θ égale à la moyenne des valeurs de θ dans c. G_i^2 est distribué comme un Khi-deux avec un nombre de degrés de liberté égal au nombre de classes.

Pour le cas où $10 < k \leq 20$, les auteurs proposent un autre test.

Les simulations de McKinley et Mills (1985) indiquent une légère supériorité de G_i^2 sur les tests de Bock et de Yen (en particulier, moins d'erreurs de type I). Un des intérêts de G_i^2 est qu'il permet facilement de comparer l'adéquation des données au modèle de Rasch et au modèle à deux paramètres (cf. Waller, 1981).

3.3 .Tests de l'invariance des paramètres

Une autre manière de tester l'adéquation au modèle consiste à éprouver l'hypothèse de l'invariance des paramètres d'items par rapport aux sous-échantillons de sujets qui servent à leur estimation. Il suffit donc d'effectuer une bipartition de l'échantillon, d'estimer les paramètres d'items dans chaque sous-échantillon et de comparer les valeurs obtenues, dont la différence doit être contenues dans les limites des fluctuations aléatoires.

A) Méthodes graphiques

20°. Comparaison graphique classique des paramètres d'items

Une méthode simple et classique pour éprouver l'hypothèse d'invariance consiste à mettre en correspondance la valeur de b de chaque item dans un sous-échantillon avec sa valeur dans l'autre. Si les données sont parfaitement conformes au modèle, les points doivent être alignés sur une droite passant par l'origine. La valeur de l'ajustement à cette droite peut être mesurée par un coefficient de corrélation de Bravais-Pearson.

Cette méthode permet de juger l'équation globale et de repérer les items nettement non conformes, mais elle manque de précision. Une amélioration simple, indiquée par Hambleton et Swaminathan (1985), consiste à subdiviser *aléatoirement* chacun des deux sous-échantillons A et B. On dispose ainsi de quatre groupes A1, A2, B1 et B2. On estime les paramètres d'item dans chacun de ces groupes et on en compare graphiquement les valeurs dans A1 et A2, puis dans A1 et B1 ; dans B1 et B2, puis A2 et B2. Les différences entre les sous-échantillons A et B sont ainsi être confrontées aux variations aléatoires.

21°. Bipartition selon le score à un item : méthode de Molenaar (1983)

Molenaar (1983) reprend l'idée de van den Wollenberg (1982, a) d'une bipartition de l'échantillon en fonction du score à un item quelconque (*cf* 24° *infra*). Molenaar préfère une méthode graphique à un test statistique, car elle présente l'avantage de faire apparaître les groupes d'items homogènes.

Les paramètres estimés dans chaque groupe sont portés sur un graphique. Les items mesurant la même dimension latente que le "subdiviseur" doivent être plus difficiles pour les sujets ayant le score 0 à cet item que pour ceux obtenant le score 1. Les items sans relation avec la dimension latente mesurée par le subdiviseur doivent être de difficulté égale dans les deux groupes et se répartir le long de la diagonale principale. On dispose ainsi d'un moyen de repérer des groupes d'items homogènes.

Ce test est très facile à mettre en oeuvre avec PML-PC. On trouvera dans l'article de 1983 un exemple traité de façon détaillée (p. 50-55) et des informations indispensables pour interpréter correctement les graphiques. Pour éviter les aléas, il convient de faire plusieurs analyses mettant en jeu des subdiviseurs différents. La méthode semble surtout convenir pour une analyse exploratoire des données.

B) Tests statistiques

22°. Test de Fischer-Scheiblechner (1970)

Le test de Fischer et Scheiblechner (1970) compare la difficulté d'un item calculé sur deux sous-échantillons différents A et B (constitués classiquement par partage à la médiane de r, mais n'importe quelle autre bipartion est possible). Pour un item, la formule du test est :

$$S_i = (b_{Ai} - b_{Bi}) / [\, v(b_{Ai}) + v(b_{Bi})]^{1/2} \qquad [32]$$

où b_{Ai} et b_{Bi} sont les estimateurs du paramètre de difficulté dans A et B, dont $v(b_{Ai})$ et $v(b_{Bi})$ sont les variances. S_i est parfois appelé t, mais Fischer (1974) précise qu'il ne s'agit en aucun cas d'un t de Student. S_i suit asympotiquement la loi normale N. Le signe de S_i peut fournir une indication sur la cause de l'inadéquation : un signe négatif suggère que l'item a une discrimination trop faible, un signe positif, qu'il a une discrimination plus forte que la moyenne.

On obtient un test global par sommation des S_i^2 :

$$S = \Sigma\, S_i^2 \qquad [33],$$

qui devrait en principe se distribuer comme un Khi-deux à (k-1) degrés de liberté. Mais Fischer (1974) indique lui-même que S ne peut suivre cette la loi que si les estimations des paramètres sont indépendantes, ce qui n'est pas le cas. Une étude de simulation de van den Wollenberg (1982, b) confirme que S n'est pas distribué comme un Khi-deux. Les incertitudes quand à la distribution de cette statistique incitent à ne pas utiliser le test de Fischer-Scheiblechner comme un test global (S), mais plutôt comme test analytique (S_i). Le paragraphe qui suit présente un autre élément de discussion.

23°. Test général de comparaison de paramètres de Lord (1980)

Lord (1980) établit une formule générale pour la comparaison des paramètres d'un item dans deux échantillons A et B. Il est généralement présenté comme un test de biais d'item (cf. § 1.4). Rien ne s'oppose cependant à l'utiliser comme test d'adéquation. Sa formule est :

$$\chi 2 = \mathbf{v}' \mathbf{M}^{-1} \mathbf{v} , \qquad [34]$$

où **v** est le vecteur des différences entre les estimateurs des p paramètres de l'item et où $\mathbf{M}^{-1}$ est l'inverse la matrice de variances-covariances de ces différences (**M** est la somme des deux matrices $\mathbf{M}_A$ et $\mathbf{M}_B$ de variances-covariances des estimateurs des paramètres). Le Khi-deux de Lord a p degrés de liberté. Dans le cas du modèle à un paramètre, la formule est simplifiée :

$$\chi_1 2 = (b_A - b_B)^2 / v(b_A) + v(b_B) \qquad [35]$$

où $v(b_1)$ et $v(b_2)$ sont les variances de b_A et de b_B .

La distribution théorique de la statistique de Lord suppose que θ est connu et non pas estimé. Dans une étude de simulation McLaughin & Drasgow (1987) constatent que le test fonctionne bien quand la localisation des sujets est connue et celle des items estimée, mais qu'il entraîne de nombreuses erreurs de type I quand la localisation des sujets et des items est estimée simultanément par le JML. Une recherche récente de Cohen et Kim (1993), où les paramètres sont estimés par le MML, conclut à la fiabilité du test : les erreurs des deux types sont peu fréquentes. Il semble donc que le test de Lord soit assez sûr quand les paramètres incidents ne sont pas estimés en même temps que les paramètres structuraux (cf. § 1.5).

En se reportant au paragraphe précédent, on peut constater que la statistique s_i^2 de Fischer-Scheiblechner et celle de Lord sont formellement identiques. Cette identité, non signalée dans la littérature, devrait conduire à prêter au test de Fischer-Scheiblechner les propriétés du test de Lord.

24°. Bipartition selon le score à un item (Van den Wollenberg (1982,a)

Van den Wollenberg (1982, a) propose de subdiviser l'échantillon, non pas sur la base du score total, mais sur celle du score à un item quelconque j.

La justification théorique du test peut être présentée ainsi. Supposons que l'échelle mesure deux dimensions latentes indépendantes A et B et que j mesure A. On peut s'attendre à ce que les items mesurant A aient une difficulté moyenne inférieure à ceux mesurant B dans le sous-échantillon des sujets qui ont réussi j ; au contraire, dans le sous-échantillon ayant échoué à j, les items mesurant A doivent avoir une difficulté moyenne supérieure à ceux mesurant B. Supposons à présent que j mesure B ; le pattern inverse doit s'observer. Par conséquent, que j mesure une dimension ou l'autre n'a pas d'importance : dans les deux cas les paramètres d'items n'ont pas la même valeur dans les deux sous-échantillons. On peut donc choisir un item au hasard, opérer une bipartition de l'échantillon selon le score des sujets à cet item et estimer les paramètres des *autres* items dans les deux sous-échantillons. Si l'échelle n'est pas unidimensionnelle, les paramètres d'items doivent avoir des valeurs différentes dans les deux groupes. Le raisonnement peut être généralisé à une échelle mesurant plus de deux dimensions latentes.

Van den Wollenberg montre au moyen d'une étude de simulation que le même test d'adéquation (ici, le test Q_1 de l'auteur) échoue à détecter la bidimensionnalité d'une échelle quand on utilise le score global comme critère de bipartition, alors qu'il met en évidence cette bidimensionnalité de façon très nette quand on se sert du score à un item pris au hasard.

Van den Wollenberg estime que ce test est moins puissant que le test de deuxième ordre Q_2, mais il le préconise néanmoins en raison de sa simplicité. Une de ses limites est que l'item choisi au hasard peut être bidimensionnel. Il est donc recommandé d'utiliser successivement plusieurs items. Une autre de ses limites, sans doute la plus importante, est qu'il est uniquement un test d'unidimensionnalité. Si les données ne sont pas conformes au modèle pour

une autre raison (discrimination différente par exemple), l'inadéquation risque de ne pas être mise en évidence par le test.

CONCLUSION

Malgré les faibles différences qui séparent certains tests et les restrictions d'emploi de ceux qui sont liés à une méthode d'estimation particulière, les tests d'adéquation au modèle de Rasch sont incontestablement nombreux. Mais aucun n'est totalement fiable. De plus, ils n'ont pas tous les mêmes propriétés : les formes d'inadéquation auxquelles ils sont sensibles ne sont pas toujours les mêmes et les informations qu'ils donnent sont également différentes (informations globales ou analytiques, information sur le seul degré d'inadéquation ou information sur les causes). C'est assez dire que le chercheur a tout intérêt à utiliser plusieurs méthodes. Il convient à cet égard de souligner la complémentarité des méthodes graphiques et des tests statistiques : les premières sont susceptibles de faire apparaître des informations supplémentaires et sont moins sensibles aux artefacts que peut engendrer un échantillon trop vaste (Gustafsson, 1980, b ; Whitely, 1980).

Les limites des tests d'adéquation incitent à ne pas les appliquer aveuglément. Hambleton préconise dans toutes ses publications (*v.g.* Hambleton, 1989 ; Hambleton & Murray, 1983 ; Hambleton & Swaminathan, 1985 ; Hambleton, Swaminathan & Rogers, 1991), une stratégie de l'évaluation de l'adéquation comportant deux aspects principaux. Le premier consiste à vérifier, préalablement à tout test d'adéquation, que les données satisfont les hypothèses du modèle. Cette analyse peut être conduite aussi bien sur des bases théoriques qu'empiriques (des méthodes nombreuses sont indiquées dans les publications précitées). Le deuxième aspect de la stratégie consiste à ne prendre des décisions que sur la base d'indications convergentes. Hambleton recommande non seulement d'utiliser plusieurs conjointement plusieurs tests, de type différent si possible, mais également de répliquer l'étude sur un autre échantillon.

Les insuffisances assez nombreuses que nous avons mentionnées ne doivent pas conduire à une position pessimiste ou sceptique. Il est vrai que si l'on peut prouver l'inadéquation des données au modèle de Rasch, on ne peut jamais être certain qu'elles lui sont totalement conformes : la conformité est toujours une conclusion provisoire. Mais il en va de même de n'importe quelle proposition théorique bien formée : on peut la réfuter, mais on ne peut pas la prouver autrement que par défaut. En outre, comme l'indique Gustafsson (1980, b), l'exigence de conformité dépend des objectifs poursuivis. Si le modèle sert de critère (test d'unidimensionnalité par ex.), le niveau d'exigence du chercheur doit être élevé. Il convient alors d'étudier l'adéquation sous tous ses angles, de réunir un effectif important, de choisir des tests puissants et des seuils de signification élevés. Mais si le modèle sert à résoudre des problèmes pratiques (en vue d'une mesure adaptative par ex.) le niveau d'exigence peut être moindre et l'on peut se contenter d'une conformité approximative, éprouvée par des méthodes moins sévères.

REMERCIEMENTS

Cette recherche a été financée partiellement par les crédits mis à notre disposition par le Ministère de l'Éducation Nationale (Direction de l'Évaluation et de la Prospective, appel d'offres *L'investissement éducatif et son efficacité*). Nous remercions vivement le Ministère pour sa générosité.

BIBLIOGRAPHIE

ANDRICH, D. (1988). *Rasch models for measurement.* Newbury Park (CA) : Sage Publications.

ANDERSEN, E.B. (1972, a). The numerical solution of a set of conditional estimation equations. *The Journal of the Royal Statistical Society*, Series B, *34*, 42-54.

ANDERSEN, E.B. (1973, b). Conditional inference in multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26*, 31-44.

ANDERSEN, E.B. (1973, a). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123-140.

BAKER, F.B. (1987). Methodology review : Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement, 11,* 111-141.

BISHOP, Y.M.M., FIENBERG, S.E. & HOLLAND, P.W. (1975). *Discrete multivariate analysis : Theory and practice.* Cambridge, MA : MIT Press.

BONIS, M. de, FÉLINE, A., LEBEAUX, M.-O. & SIMON, M. (1994). Evaluation de la sévérité de la dépression : Comparaison des modèles logistique, factoriel et implicite. In M. Huteau (Ed.), *Les techniques d'évaluation des personnes* (68-70). Paris : E.A.P.

BOCK, R.D. & AITKIN, M. (1981). Marginal maximum likehood estimation of item parameters : An application of an EM algorithm. *Psychometrika, 46*, 443-459.

COHEN A. S. & KIM, S.-H. (1993). A comparison of Lord's χ^2 and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17*, 39-52.

DICKES, P. (1983). Modèle de Rasch pour items dichotomiques : Théorie, technique et application à la mesure de la pauvreté. *Cahiers Economiques de Nancy, 11*, 73-116.

DICKES, P. & HAUSMAN, P. (1983). Définir et mesurer la délinquance juvénile. *Bulletin de Psychologie,* n° 359, 441-455.

DIVGI D. R. (1981). Model free evaluation of equating and scaling. *Applied Psychological Measurement, 5*, 203-208.

ELLIOT, C. D., MURRAY, D. J. & SAUNDERS, R. (1977). *Goodness of fit to the Rasch model as a criterion of unidimensionality.* Manchester : University of Manchester.

FLIELLER, A. (1988). Application du modèle de Rasch à un problème de comparaison de générations. *Bulletin de Psychologie, 42,* 86-91.

FRASER, C. & McDONALD, R.P. (1988). NORHAM : Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267-269.

FISCHER, G. H. (1974). *Einführung in die Theorie psychologischer Tests.* Bern : Verlag Hans Huber.

FISCHER, G. H. & SCHLEIBLECHNER, H.H. (1970). Algorithmen und Programmen für das probabilistische Testmodel von Rasch. *Psychologische Beiträge, 12*, 23-51.

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

GOLDSTEIN, H. & WOOD, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42*, 139-167.

GUSTAFSSON, J.E. (1980, a). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement, 40*, 377-385.

GUSTAFSSON, J.E. (1980, b). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 32*, 205-233.

GUSTAFSSON, J.E. & LINBLAD, T. (1978). The Rasch model for dichotomous items : A solution of the conditional estimation problem for long tests and thome thoughts about item screening procedures. Reports from the Institute of Education, University of Göteborg, n° 67.

HAMBLETON, R.K. (1969). *An empirical investigation of the Rasch test theory model.* Unpublished doctoral dissertation, University of Toronto (Canada).

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

HAMBLETON, R.K. & MURRAY, L. (1983). Some goodness of fit investigations for item response models, in R. K. Hambleton (Ed.), *Applications of item response theory* (p. 71-94).Vancouver : Educational Research Institute of British Columbia.

HAMBLETON, R.K. & SWAMINATHAN, H. (1985). *Item Response Theory*. Boston (MA) : Kluver-Nihoff Publishing.

HAMBLETON, R.K. & SWAMINATHAN, H. & ROGERS, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park (CA) : Sage Publications.

HATTIE, J. (1985). Methodology review : Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

LUDLOW, L.H. (1985). A strategy for the graphical representation of Rasch model residuals. *Educational and Psychological Measurement, 45*, 851-859.

LUDLOW, L.H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement, 10*, 217-229.

LORD, F. (1980). *Applications of Item Response Theory to practical testing problems.* Hillsdale (N.J.) : Lawrence Erlbaum.

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

McKINLEY, R. L. & MILLS C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49-57.

McLAUGHLIN, M. E. & DRASGOW, F. (1987). Lord's Chi-Square Test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11*, 161-173.

MEAD, R. (1976). *Assessment of fit data to the Rasch model.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

MARTIN-LÖF (1973, oktober). *Statistika Modeller.* Anteckningar från seminarier Läsaret 1969-1970 utarbetabe av Rolf Sunberg. Stockholm : Instituet för Försäkringsmathematik och Matematisk Statistik vid Stockholms Universitet.

MISLEVY, R.J. & BOCK, R.D. (1990). *Bilog 3* (2nd ed). Mooresville (IN) : Scientific Software Inc.

MOLENAAR, I.W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika, 48*, 49-72.

MOLENAAR, I.W. (1990). *P.M.L. : User's manual PC version.* Groningen : ProGAMMA.

RASCH, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen : Danish Institute for Educational Research (1st ed.) / Chicago : University Press (2nd ed.).

RASCH, G. (1977). On specific objectivity : an attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14*, 58-94.

REISER, M. (1989). An application of the item-response model to psychiatric epidemiology. *Sociological Methods and Research, 18*, 66-103.

REUCHLIN, M. (1992). *Introduction à la recherche en psychologie.* Paris : Nathan.

ROGERS, H. J. & HATTIE, J. A. (1987). A Monte Carlo investigation of several person and item fit statitistics for item response models. *Applied Psychological Measurement, 11*, 47-57.

STENE E. (1969). *An exact test for stochastic independence of responses in an item analysis model.* Symposium on Rasch models, Køge (Denmark).

STOCKING, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika, 55*, 461-475.

SWAMINATHAN, H. & GIFFORD, J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*, 175-191.

THISSEN, D. (1982). Marginal maximum likehood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175-186.

TRAUB, R. E. & LAM, R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology, 36*, 19-48.

Van den VIJVEN, F.R. (1986). The robustness of Rasch estimates. *Applied Psychological Measurement, 10,* 45-57.

Van den WOLLENBERG, A. L. (1979). *The Rasch model and time limit tests.* Doctoral dissertation. University of Nijmegen (The Netherlands) : Student papers.

Van den WOLLENBERG, A. L. (1982, a). A simple and effective method to test the dimensionality axiom of the Rasch model. *Applied Psychological Measurement, 6,* 83-91.

Van den WOLLENBERG, A. L. (1982, b). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-140.

WALLER, M.I. (1981). A procedure for comparing latent trait models. *Journal of Educational Measurement, 18*, 119-125.

WHITELY, S.E. (1980). Latent trait models in the study of intelligence. *Intelligence, 4*, 97-132.

WRIGHT, B. D. (1985). Rasch measurement models. In T.H. Husén & T.N. Postlethwaite (Eds). *The International Encyclopedia of Education*, 1st ed. (4177-4181). Oxford : Pergamon Press.

WRIGHT, B. D. & DOUGLAS, G.A. (1977). Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement, 37*, 573-586.

WRIGHT, B. D., MEAD, R. J. & BELL, S. R. (1979). *BICAL : Calibrating items with the Rasch model.* Statistical Research Memorandum N° 23B. Chicago : University of Chicago, School of Education.

WRIGHT, B.D. & PANCHAPAKESAN, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-57.

WRIGHT, B.D. & STONE, M.H. (1979). *Best test Design.* Chicago : MESA Press.

YEN, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

YEN W. M. (1987).A comparison of the efficiency and accuracy of bilog and logist. *Psychometrika, 52*, 275-291.