Modèles de Réponse à l'Item pour items polytomiques : exemple d'utilisation du logiciel MULTILOG

Géraldine Rouxel

Université Rennes 2 - CRPCC, Laboratoire de psychologie expérimentale, Groupe de recherche en psychologie différentielle - 6 avenue Gaston Berger - 35043 Rennes cedex (Mel: geraldine.rouxel@uhb.fr).

Psychologie et Psychométrie, 113-130, vol.20, n°2/3, 1999

Summary

After different item response models for polytomous items were reviewed, the two parameters Samejima's Graded Response Model was described more precisely. MULTILOG was then used to apply this model to the study of a personality scale (task involvement measure) completed by 505 students (4^{th} and 5^{th} grades). The analysis of the « discrimination » and « threshold » parameters, as well as the item characteristic curves, item information curves and test information curves allowed a better understanding of the part played by each item within the scale. Those results were in accordance with those obtained within the framework of a classical reliability analysis. Nevertheless, we conclude that the greatest care must be taken in using this kind of model by rule of thumb, in particular when applied to the study of personality scales.

Key words

Item Response Models - Polytomous items - MULTILOG - Personality scale

Résumé

Après avoir passé rapidement en revue différents Modèles de Réponse à l'Item pour items à réponses polytomiques, nous nous sommes attachée à la description plus précise du modèle à deux paramètres des Réponses Graduées de Samejima. Grâce à l'utilisation du logiciel MULTILOG, nous avons alors appliqué ce modèle à l'étude d'un questionnaire de personnalité (mesure de l'orientation de but vers la tâche) complété par 505 élèves de CM1-CM2. L'analyse des paramètres « discrimination » et « seuil », ainsi que celle des courbes

caractéristiques des catégories de réponse de chaque item, des fonctions d'information des items et de la courbe d'information de l'échelle ont permis une meilleure compréhension du fonctionnement de chacun des items au sein du questionnaire. Les résultats ainsi dégagés ont de plus été confortés par ceux obtenus dans le cadre d'une analyse d'items classique. On conclut néanmoins en incitant à la plus grande prudence quant à une utilisation qui tendrait à devenir routinière de ce type de modèle, en particulier dans le cas de l'étude de questionnaires de personnalité.

Mots Clés

Modèles de Réponse à l'Item - Items polytomiques - MULTILOG - Echelle de personnalité

INTRODUCTION

Dans le cadre de la théorie classique des tests, l'aptitude d'un individu est définie en référence à un test particulier : le niveau d'aptitude dépend du niveau de difficulté du test. Se pose alors inévitablement la question de la définition de la difficulté d'un test. Question à laquelle on répond en estimant la proportion de sujets qui, dans un groupe déterminé, répond correctement aux items du test considéré. Ainsi, le degré de difficulté d'un test dépend-il de l'aptitude des sujets étudiés, alors que l'aptitude des sujets dépend de la difficulté des items du test auquel ils ont été soumis... De la même façon, le pouvoir discriminant des items, la fidélité et la validité des scores à un test sont définis en référence à un groupe particulier d'individus. Les caractéristiques des items et des tests changeront donc avec la population étudiée ; de même que les caractéristiques attribuées aux sujets évolueront lors de l'utilisation de tests différents. Il est par conséquent très difficile à la fois de comparer des sujets qui passent des tests différents et des items dont les caractéristiques sont obtenues en utilisant des groupes disparates de sujets. En outre, la théorie classique des tests ne nous permet pas de savoir quelle est la probabilité pour un individu particulier de répondre correctement à un item donné (Weiss & Yoes, 1991). En résumé, on pourrait souhaiter qu'une théorie des tests puisse au moins présenter les propriétés suivantes (Hambleton, Swaminathan & Rogers, 1991):

a) fournir des caractéristiques d'items indépendantes du groupe à partir duquel elles sont obtenues,

b) permettre le calcul de scores décrivant l'aptitude des sujets qui soient indépendants du test utilisé,

c) s'appuyer sur un modèle s'exprimant au niveau de l'item plutôt que du test,

d) renvoyer à un modèle ne requérant pas strictement des tests parallèles pour évaluer la fidélité,

e) proposer un modèle fournissant une mesure de précision pour chaque score d'aptitude.

Dans le cadre d'une autre théorie des tests, les Modèles de Réponse à l'Item (MRI) permettent de répondre à ces exigences. Plusieurs MRI pour items dichotomiques existent et sont décrits ailleurs dans ce numéro, nous nous limiterons donc à la présentation de modèles pour données polytomiques.

MODELES POUR ITEMS A REPONSES POLYTOMIQUES

Généralités

Thissen et Steinberg (1986) ont proposé une taxonomie permettant de classer les divers modèles de réponse à l'item en cinq catégories. Les MRI pour données polytomiques et ordonnées sont rassemblés dans les deux catégories qui suivent (tableau 1) (Il existe une troisième catégorie réservée aux items polytomiques, mais pour items à choix multiples. Les deux autres catégories concernent les items à réponses dichotomiques.):

Modèles de « différence »	Modèles « division par le total »
Modèle des réponses graduées (*Graded response)* (Samejima) Modèle de l'échelle de réponses (*Rating scale*) (Muraki)	Modèle des réponses nominales (*Nominal response*) (Bock) Modèle du crédit partiel généralisé (*Generalized partial credit*) (Muraki) Modèle du crédit partiel (*Partial credit**) (Masters) Modèle des intervalles successifs (*Successive intervals**) (Rost) Modèle de l'échelle de réponses (*Rating scale**)(Andrich)

Tableau 1 - Deux catégories de MRI polytomiques d'après la taxonomie proposée par Thissen et Steinberg (1986) (* Modèles appartenant à la famille des modèles de Rasch)

Dans les modèles dits de « différence », la probabilité de répondre en choisissant telle ou telle catégorie est calculée en soustrayant la probabilité de répondre dans une catégorie donnée ou plus élevée (conditionnelle au niveau du trait, θ) de la probabilité de répondre dans la catégorie adjacente ou plus faible (conditionnelle à θ) : par exemple, la probabilité de répondre dans la catégorie k est $P^*(k) - P^*(k+1)$. L'équation décrivant la probabilité de réponse dans chacune des catégories est appelée « fonction caractéristique opératoire » du modèle (*Operative Characteristic Function*). Le modèle que l'on présentera par la suite plus précisément, le Modèle des Réponses Graduées (MRG) de Samejima, appartient à cette première catégorie.

Avec les modèles « division par le total », contrairement aux modèles de « différence », on peut obtenir directement la fonction caractéristique opératoire. En effet, la probabilité de répondre dans une catégorie donnée est calculée en divisant le numérateur par la somme de tous les numérateurs de chacune des probabilités associées à une catégorie, de façon à ce que la somme des probabilités conditionnelles à θ soit égale à 1. Le modèle du crédit partiel et le modèle de l'échelle de réponses (voir Andrich, 1995 ; De Ayala, 1993), qui appartiennent à cette deuxième catégorie, sont des modèles fréquemment utilisés. Le modèle du crédit partiel (proposé par Masters), comme le MRG de Samejima, s'applique à des items dont les catégories de réponse sont ordonnées. Mais il diffère de ce dernier en ce qu'il appartient à la famille des modèles de Rasch et en partage donc les propriétés. Le modèle du crédit partiel est le plus simple de tous les MRI pour catégories ordonnées. Comme le modèle de Rasch pour items dichotomiques, il ne comprend en effet que deux ensembles de paramètres : les paramètres « aptitude » et « difficulté » (Masters & Wright, 1997). A chaque score brut total qu'il est possible de calculer à partir des scores attribués à chacune des réponses aux items d'un questionnaire correspond un paramètre « aptitude ». Le nombre de paramètres « difficulté » (on parle aussi de « seuils » ou de points sur le continuum latent correspondant aux croisements des courbes caractéristiques de chacune des catégories de réponse) dépend du nombre de catégories de réponse : on en calcule un entre deux catégories adjacentes. Le modèle de l'échelle de réponses (proposé par Andrich) est très proche du modèle du crédit partiel (il appartient aussi à la famille des modèles de Rasch), il s'applique également lorsque les catégories de réponse sont ordonnées, mais contrairement à ce dernier il suppose en plus un égal espacement entre deux catégories de réponse adjacentes pour tous les items d'un questionnaire (voir Andersen, 1997). Ainsi, est-il possible de calculer la valeur du seuil entre les catégories 2 et 3, en ajoutant à la valeur de seuil entre les catégories 1 et 2 une même constante pour tous les items. De même, la valeur du seuil entre les catégories 3 et 4 s'obtient-elle en ajoutant une même constante (mais qui peut être différente de celle permettant le passage du seuil 1 au seuil 2) pour tous les items, à la valeur du seuil entre les catégories 2 et 3.

Présentation du Modèle des Réponses Graduées de Samejima

Le modèle de Samejima est un MRI à deux paramètres (« difficulté » et « discrimination »), permettant d'analyser des items polytomiques, c'est-à-dire des items qui comprennent plusieurs catégories de réponses ordonnées (ce modèle date de 1969, voir Koch, 1983 ;

Thissen & Steinberg, 1988 ; Samejima, 1997). Le MRG est en fait une généralisation de l'équation (1) relative aux modèles logistiques à deux paramètres pour items dichotomiques adaptée à des items polytomiques.

$$P(\theta_i) = \frac{1}{1+e^{-a_j(\theta_i - b_j)}} \qquad (1)$$

avec θ_i le paramètre « aptitude » du sujet *i*; a_j le paramètre « discrimination » de l'item *j* ; b_j le paramètre « difficulté » de l'item *j* ; $P|\theta$ la probabilité conditionnelle de donner une certaine réponse à un item donné.

Contrairement aux modèles de la famille de Rasch, on ne peut pas avoir ici de certitudes quant à l'unicité des estimations d'un paramètre (Thissen & Steinberg, 1997). Le modèle proposé par Samejima est basé sur une fonction logistique et sur la probabilité pour que la réponse à un item soit observée dans « une catégorie *k* ou une catégorie plus élevée ». L'utilisation du MRG est pertinente quand les réponses à un item peuvent être classées dans plus de deux catégories ordonnées représentant des degrés variés d'atteinte d'une solution à un problème ou d'accord avec une proposition d'attitude. Ce modèle repose sur l'équation suivante :

$$P(x = k \mid \theta) = \frac{1}{1+e^{-a(\theta - b_{k-1})}} - \frac{1}{1+e^{-a(\theta - b_k)}} \qquad (2)$$
$$= P^*(k) - P^*(k+1)$$

L'équation (2) spécifie pour chaque individu la probabilité conditionnelle de réponse à une catégorie particulière *k* (avec *k* = 1, 2, ...m, où la réponse m renvoie à la valeur θ la plus élevée). P*(*k*) renvoie à la fonction de réponse à l'item décrivant la probabilité pour qu'un sujet réponde dans la catégorie *k* ou plus, pour chaque valeur de θ. Par définition, la probabilité de répondre au moins dans la catégorie *k* est égale à 1 [P*(1) = 1] et la probabilité de répondre dans la catégorie (*m*+1) est nulle [P*(*m*+1) = 0].

Ce modèle nécessite l'estimation d'un paramètre *a* et de (*m*-1) paramètres *b* pour chacun des items du test. Parce qu'il y a (*m*-1) seuils entre *m* catégories de réponses ordonnées, (*m*-1) fonctions de réponse sont nécessaires pour décrire les réponses à un item donné. Le paramètre *a* est le paramètre « discrimination » de l'item et sa valeur est proportionnelle à la pente des fonctions de réponse. Les paramètres de discrimination sont constants pour chacune des fonctions de réponse propre à chaque item. Le paramètre *a* par contre peut varier entre items. Les paramètres b_k sont appelés « difficulté de catégorie » ou « seuils » et sont définis par le

point sur l'échelle θ où la probabilité est de 50% pour que la réponse à un item se fasse au moins dans la catégorie *k*. Avec *k*=*m* catégories, il y a (*m-1*) seuils entre les *m* catégories de réponses ordonnées : on estime donc (*m-1*) paramètres *b* pour chaque item :

Catégories de réponses → 1 2 3 4

Seuils → b_1 b_2 b_3

Le modèle fournit un indice d'ajustement aux données : un rapport de vraisemblance (*Likelihood Ratio*, G^2) associé à une probabilité de rejet de l'hypothèse nulle d'ajustement du modèle aux données. Sous certaines conditions, G^2 est distribué comme un chi-deux avec un nombre de degrés de liberté égal au nombre de patterns de réponses moins le nombre d'estimations faites par le modèle. G^2 reflète le manque de congruence entre la fréquence des patterns de réponses observés et la fréquence de ces patterns prédits par les fonctions de réponse à l'item estimées : plus faible est cette congruence et plus forte est la valeur G^2. Ce type d'indice cependant est à considérer avec prudence, car comme c'est également le cas dans la modélisation structurale, le non rejet de l'hypothèse nulle a plus de chances de survenir lorsque le nombre de sujets est peu élevé, ce qui pousse d'ailleurs certains auteurs à douter de l'intérêt de l'utilisation de ces tests statistiques (e.g., Reuchlin, 1997). Enfin, avec des items à réponses polytomiques ou lorsque le nombre d'items est important, G^2 n'est généralement pas approprié pour juger de l'ajustement d'un modèle de base (c'est-à-dire un modèle dans lequel tous les paramètres sont librement estimés). En effet, on aboutit dans ces cas à un nombre de patterns de réponses non observés parfois considérable (par exemple, dans le cas d'un questionnaire comprenant 12 items pour lesquels les réponses se font sur une échelle likert en 4 points, 412 patterns de réponses différents sont envisageables !). La statistique G^2 suit alors une distribution de référence inconnue. Les valeurs G^2 ne peuvent être alors utilisées que dans le cadre de la comparaison de modèles emboîtés. Un autre moyen de s'assurer de l'ajustement d'un modèle (mais il en existe d'autres) consiste à vérifier l'invariance des paramètres « aptitude » (Hambleton *et al.*, 1991 ; Van der Linden & Hambleton, 1997). En effet, si l'ajustement du modèle est acceptable, les estimations « d'aptitude » des sujets doivent être à peu près les mêmes (aux erreurs de mesure près) dans des échantillons différents d'items tirés d'un même test ou questionnaire (par exemple, les items « faciles » versus « difficiles »). On rappelle en effet que dans ces modèles, par définition, les paramètres caractérisant les sujets ne doivent pas dépendre des items.

EXEMPLE D'APPLICATION EMPIRIQUE DU MODELE DES REPONSES GRADUEES DE SAMEJIMA

Présentation de l'étude

Pour illustrer notre propos on se base sur les données recueillies auprès de 505 élèves en classes de CM1-CM2 (âge moyen : 10;4 ans) grâce à un questionnaire d'auto-évaluation destiné à mesurer l'orientation de but vers la tâche dans le domaine du français. Il s'agit d'une orientation motivationnelle qui pousse l'individu à se focaliser sur le développement de savoir-faire et de compétences relatifs à la tâche qu'il doit exécuter. Ce questionnaire est une traduction du « *Goal-Orientation Questionnaire* » construit par Seegers et Boekaerts (1993) (tableau 2) :

1- Je suis content quand j'ai appris quelque chose en français qui a du sens pour moi.
2- Quand je ne réussis vraiment pas à terminer un exercice de français, je demande de l'aide.
3- Quand je n'ai pas réussi un exercice en français, je ne suis pas content de moi.
4- Je préfère des exercices de français difficiles qui m'apprennent des choses nouvelles, plutôt que des exercices faciles.
5- Quand j'ai une plus mauvaise note que d'habitude en français, je suis déçu.
6- J'aime réfléchir à un exercice de français jusqu'à ce que je trouve la réponse.
7- Je ne suis pas content de moi quand je n'ai pas travaillé assez dur en français.
8- Je suis content quand j'ai appris quelque chose d'intéressant en français.
9- Quand je n'ai pas fait mes exercices de français aussi bien que d'habitude, je ne suis pas content de moi.
10- J'aime bien quand j'apprends quelque chose de nouveau en français.
11- Quand je ne vois pas immédiatement comment faire un exercice en français, je fais encore plus d'efforts.
12- Je ne suis pas content avec moi quand il y a quelque chose que je n'ai pas compris pendant la leçon de français.

Tableau 2 - Items du questionnaire « Orientation de but vers la tâche »

On propose aux sujets quatre catégories de réponses, ordonnées, allant de la proposition « presque jamais vrai pour moi » à « presque toujours vrai pour moi », en passant par « parfois vrai pour moi » et « souvent vrai pour moi ». On précise que lorsque dans un pattern de réponses individuel le nombre de données manquantes est strictement inférieur à la moitié du nombre de réponses possibles, on remplace ces valeurs manquantes par la valeur la plus fréquente donnée pour cette réponse au sein de l'échantillon. Sinon, quand le nombre de données manquantes devient trop important, le pattern en question est retiré de l'analyse.

Analyse d'items classique

L'analyse d'items fournit un α de Cronbach égal à .82 traduisant une consistance interne de l'échelle satisfaisante. Au niveau de l'item, on a calculé les indices de « difficulté » (moyenne calculée pour chaque item) et de « discrimination » de chacun des items (corrélations item-échelle). L'ensemble des résultats de l'analyse figure dans le tableau ci-dessous (tableau 3) :

	Moyenne	Coefficient de variation	r item-échelle	r^2	α si item éliminé
Item 1	3,02	32	.46	.28	.81
Item 2	1,97	50	.17	.06	.83
Item 3	2,43	43	.49	.31	.80
Item 4	2,71	40	.41	.27	.81
Item 5	2,81	36	.46	.30	.81
Item 6	2,67	37	.48	.30	.80
Item 7	2,42	42	.60	.42	.79
Item 8	3,05	32	.55	.41	.80
Item 9	2,49	42	.58	.42	.79
Item 10	3,05	32	.48	.32	.80
Item 11	2,77	36	.49	.27	.80
Item 12	2,40	43	.46	.29	.81

Tableau 3 - Résultats de l'analyse d'items - Questionnaire « Orientation de but vers la tâche »

Sur les douze items, les moyennes des réponses s'échelonnent entre 1,97 et 3,05 avec une moyenne globale égale à 2,65 (ces moyennes sont évidemment à considérer avec prudence étant donné qu'elles sont calculées à partir d'échelles ordinales). On ne peut manquer de remarquer l'item 2 qui présente une moyenne nettement inférieure à celle des autres: les sujets ont eu tendance à donner des réponses allant dans le sens d'un niveau plutôt faible d'orientation de but vers la tâche. On observe également pour cet item une plus grande dispersion des réponses (coefficient de variation égal à 50, la moyenne de l'ensemble de ces coefficients étant de 36,25). Concernant les indices de « discrimination » des items, on obtient une valeur moyenne de 0,47. A nouveau l'item 2 se distingue avec une valeur nettement inférieure (égale à 0,17) dénotant un très faible pouvoir discriminant. C'est enfin le seul item dont l'élimination engendrerait une légère amélioration de la consistance interne du questionnaire.

Ce type d'analyse apporte évidemment des renseignements intéressants, mais parce que les indices d'items et d'échelle qu'il fournit sont spécifiques à l'échantillon à partir duquel ils ont

été calculés, les interprétations basées sur ces statistiques sont difficilement généralisables. C'est à ce niveau que l'emploi des MRI peut s'avérer très intéressant. De plus, une des limites de l'analyse d'items classique renvoie au fait qu'elle nous permet de juger de la consistance interne d'un questionnaire, mais pas de son homogénéité (unidimensionnalité). La consistance interne est une condition nécessaire à l'homogénéité, mais pas suffisante (Schmitt, 1996). Les MRI permettent par contre de juger de l'homogénéité d'une échelle.

Analyse des items avec le Modèle des Réponses Graduées de Samejima

Le logiciel MULTILOG (Thissen, 1991) offre la possibilité d'estimer les paramètres d'items du MRG de Samejima. Dans le cas du questionnaire étudié en exemple, le modèle estime un paramètre « discrimination » et trois paramètres « difficulté » pour chaque item (c'est-à-dire dans cet exemple, leur capacité à engendrer un niveau élevé d'orientation de but vers la tâche). MULTILOG permet l'estimation des paramètres soit grâce à la méthode d'estimation par maximum de vraisemblance (*Maximum Likelihood*), soit par la méthode du maximum de vraisemblance marginale (*Marginal Maximum Likelihood*).

Etant donné que nous ne pouvions nous fier à la valeur G^2 pour connaître l'ajustement de notre modèle (le nombre de patterns observés étant très inférieur au nombre de patterns possibles), nous avons procédé à la vérification de l'invariance des paramètres « aptitude » estimés par le modèle. Nous avons ainsi partagé les items en deux groupes égaux (les patterns de réponses possibles dans ces deux groupes doivent être identiques) selon non pas un critère de difficulté qui n'aurait pas de sens ici, mais en fonction de leur tendance à engendrer un plus ou moins grand nombre de réponses dans les catégories 1 et 2 d'une part, 3 et 4 d'autre part. Le modèle de Samejima est ensuite testé dans les deux sous-ensembles d'items. On repère alors les patterns de réponses communs aux deux analyses auxquels on va attribuer, d'une part, les paramètres « aptitude » estimés à partir des réponses aux items données par le premier groupe, d'autre part, les paramètres « aptitude » estimés grâce aux réponses aux items au sein du deuxième groupe. On calcule enfin la corrélation entre ces deux distributions de scores théoriques. Si celle-ci est élevée, on vérifie l'invariance des paramètres estimés. C'est le cas pour le questionnaire étudié, on obtient en effet une corrélation très forte de .987 (n=102).

Les résultats de l'analyse effectuée (méthode du maximum de vraisemblance marginale) figurent dans le tableau ci-dessous (tableau 4) :

	Discrimination *a* (pente)	Seuil b1 (catégorie 1 à 2)	Seuil b2 (catégorie 2 à 3)	Seuil b3 (catégorie 3 à 4)
Item 1	1,24 (0,14)	-2,33 (0,30)	-0,89 (0,14)	0,47 (0,13)
Item 2	0,33 (0,09)	-1,27 (0,61)	3,22 (1,10)	6,53 (2,13)
Item 3	1,38 (0,15)	-1,27 (0,16)	0,30 (0,11)	1,23 (0,16)
Item 4	1,01 (0,12)	-1,80 (0,28)	-0,40 (0,14)	0,98 (0,19)
Item 5	1,24 (0,14)	-2,10 (0,26)	-0,32 (0,12)	0,73 (0,14)
Item 6	1,26 (0,10)	-1,83 (0,22)	-0,20 (0,12)	1,16 (0,16)
Item 7	1,90 (0,16)	-1,11 (0,11)	0,28 (0,09)	1,13 (0,12)
Item 8	1,54 (0,16)	-2,03 (0,23)	-0,87 (0,12)	0,30 (0,10)
Item 9	1,71 (0,16)	-1,15 (0,13)	0,12 (0,09)	1,07 (0,13)
Item 10	1,23 (0,14)	-2,32 (0,29)	-0,96 (0,15)	0,39 (0,12)
Item 11	1,36 (0,15)	-1,83 (0,22)	-0,36 (0,11)	0,83 (0,14)
Item 12	1,19 (0,14)	-1,30 (0,18)	0,30 (0,12)	1,45 (0,20)

Tableau 4 - Paramètres «discrimination » et « seuil » du modèle de Samejima (erreurs-standard entre parenthèses) - Questionnaire « Orientation de but vers la tâche en français »

Avec les MRI, on considère que les valeurs de discrimination inférieures à .80 sont inacceptables et celles comprises entre .80 et 1 sont tout juste acceptables. Concernant les paramètres « difficulté », on s'attend à ce que les seuils soient largement distribués de -3 à +3 le long du continuum latent, ce qui permet de différencier des sujets situés à des niveaux variés d'aptitudes (Flannery, Reise & Widaman, 1995).

Une analyse rapide de ce tableau montre que sur les 36 paramètres « seuils » calculés par le modèle, 12 (soit 1/3 d'entre eux) présentent des valeurs inférieures à -1 et seulement 7 des valeurs supérieures à +1. Cette dominance de valeurs négatives suggère que les distinctions aux niveaux « d'aptitudes » les plus élevés risquent de poser problème avec ce questionnaire (effet plancher). Par ailleurs, comme lors de l'analyse d'items précédente, l'item 2 se démarque. Il présente en effet une valeur de discrimination très faible (0,33), inacceptable selon les critères en vigueur. Cet item se caractérise également par des valeurs disproportionnées des seuils b_2 et b_3 par rapport aux autres items, ainsi que par des erreurs-standard associées anormalement élevées.

Construction des courbes caractéristiques des catégories des items

Grâce aux valeurs des différents paramètres « discrimination » et « seuil » fournies par le modèle, il est possible de construire pour chacun des items les courbes caractéristiques des catégories qui leur sont propres.

Dans le cas d'items à quatre catégories de réponses, on doit calculer les cinq probabilités suivantes (notées P*) :

a) La probabilité pour que la réponse à un item soit observée dans la catégorie 1 ou une catégorie plus élevée qui est évidemment égale à 1 :

P*(1) = P*(catégorie 1 ou plus élevée) = 1

b) La probabilité pour que la réponse d'un item soit observée dans la catégorie 2 ou une catégorie plus élevée, qui se calcule grâce à la fonction logistique utilisée pour des items à réponses dichotomiques (voir équation (1)) :

P*(2) = P*(catégorie 2 ou plus élevée)

$= \{1 + \exp [-a_j (\theta - b_{j1})]\}^{-1}$

(b_{j1} correspond au seuil entre les catégories 1 et 2 pour l'item j, a_j est le paramètre pente).

c) La probabilité pour que la réponse à un item s'observe dans la catégorie 3 ou une catégorie plus élevée :

P*(3) = P*(catégorie 3 ou plus élevée)

$= \{1 + \exp [-a_j (\theta - b_{j2})]\}^{-1}$

d) La probabilité pour observer la réponse à un item dans la catégorie 4 :

P*(4) = P*(catégorie 4)

$= \{1 + \exp [-a_j (\theta - b_{j3})]\}^{-1}$

e) Enfin, la probabilité pour que l'on observe une réponse dans une catégorie 5 ou plus élevée, qui est évidemment nulle :

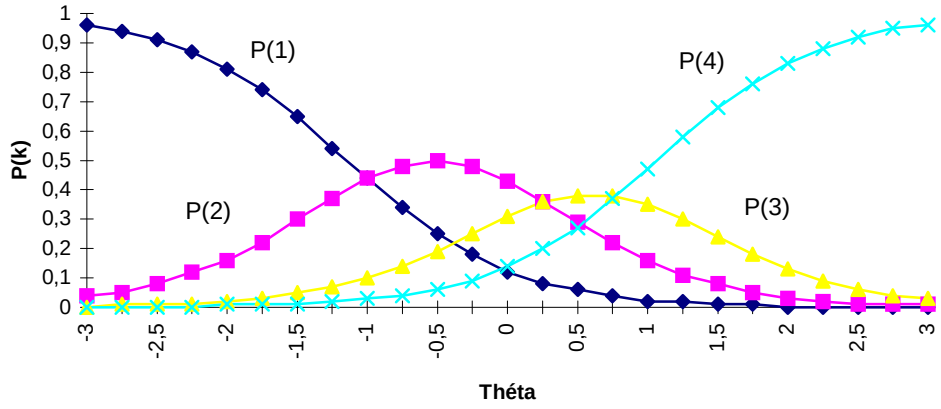
P*(5) = P*(catégorie 5 ou plus élevée) = 0

Toutes ces équations sont nécessaires pour tracer les courbes représentant la probabilité d'observer chaque réponse en fonction de θ:

P_j (catégorie k) = P* (k) - P* ($k+1$)

Ces courbes sont nommées « courbes caractéristiques des catégories d'items ». Elles s'interprètent de la même manière que celles relatives aux items à réponses dichotomiques. Les courbes caractéristiques des catégories d'items fournissent des descriptions directement interprétables des processus de réponses à la base des données, qui sous leur forme brute sont

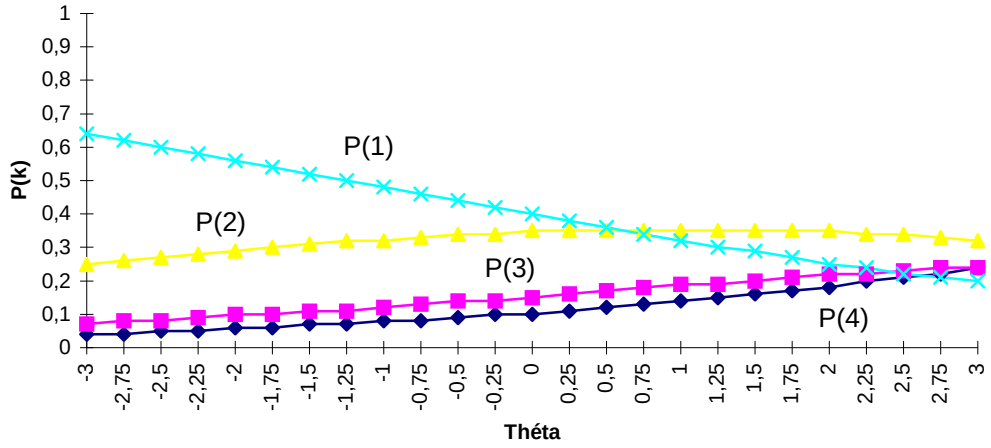
assez peu parlantes. On présente ci-dessous (graphique 1) un exemple de courbes caractéristiques des catégories d'un item au comportement conforme au modèle testé :



Graphique 1- Courbes caractéristiques des catégories de l'item 9

On rappelle que le terme « aptitude » tel qu'il est utilisé dans ce contexte est à comprendre comme renvoyant, pour ce questionnaire, à une certaine propension à privilégier une orientation motivationnelle tournée vers la tâche. P(*k*) correspond à la fonction de réponse *k* et renvoie à la probabilité pour que la réponse à l'item se trouve dans la catégorie *k* à un niveau d'aptitude θ donné. La fonction de réponse correspondant à la catégorie *k* la plus élevée (catégorie 4) tend vers 1 quand θ tend vers plus l'infini et vers 0 quand θ tend vers moins l'infini. Pour la catégorie avec le score le plus faible (catégorie 1), la fonction de réponse tend vers 0 quand θ tend vers plus l'infini et vers 1 quand θ tend vers moins l'infini. Il s'ensuit que pour les individus caractérisés par un paramètre « aptitude » élevé, la probabilité est importante de choisir la catégorie 4 ; tandis que les individus qui ont un niveau faible « d'aptitude » auront plus de chances de choisir la catégorie de réponse 1. Pour les valeurs moyennes « d'aptitudes », les probabilités de choix sont modérées pour les quatre catégories.

Pour ce questionnaire, tous les items présentent des courbes caractéristiques acceptables, relativement similaires à celles qui viennent d'être décrites. Un seul item diverge totalement ; sans surprise il s'agit de l'item 2 (graphique 2) :

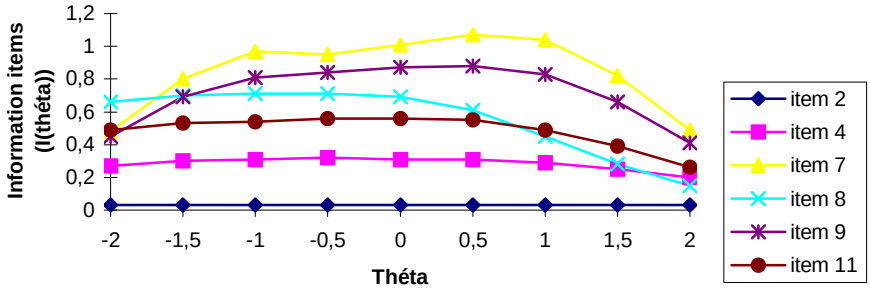


Graphique 2 - Courbes caractéristiques des catégories de l'item 2

Construction des courbes d'information d'items et d'échelle

En plus d'offrir la possibilité de décrire les items et de permettre le tracé des courbes caractéristiques des catégories d'items, l'estimation des paramètres d'item rend possible le calcul d'indices exprimant la quantité d'information apportée par chaque item et plus largement par le test. En effet, une fois que les fonctions de réponse propres à un item sont connues, c'est-à-dire que ses paramètres sont estimés, on peut les transformer en une courbe d'information d'item. Les graphiques d'information d'item représentent le niveau d'aptitude du sujet (θ) sur l'axe des abscisses et la quantité d'information apportée au questionnaire par l'item sur l'axe des ordonnées (I(θ)). L'allure d'une courbe d'information d'item est dépendante des paramètres « discrimination » et « difficulté ». Plus la discrimination est importante, plus la courbe d'information de l'item sera pointue et plus ce dernier sera informatif. Les paramètres « difficulté » déterminent la localisation des différentes courbes d'information. Les items « faciles » (c'est-à-dire ceux dont les paramètres *b* sont inférieurs à -1) fournissent de l'information dans la zone des niveaux faibles d'aptitude. Les items « difficiles » (c'est-à-dire ceux dont le paramètre *b* est supérieur à 1) fournissent de l'information dans la région des niveaux élevés d'aptitude.

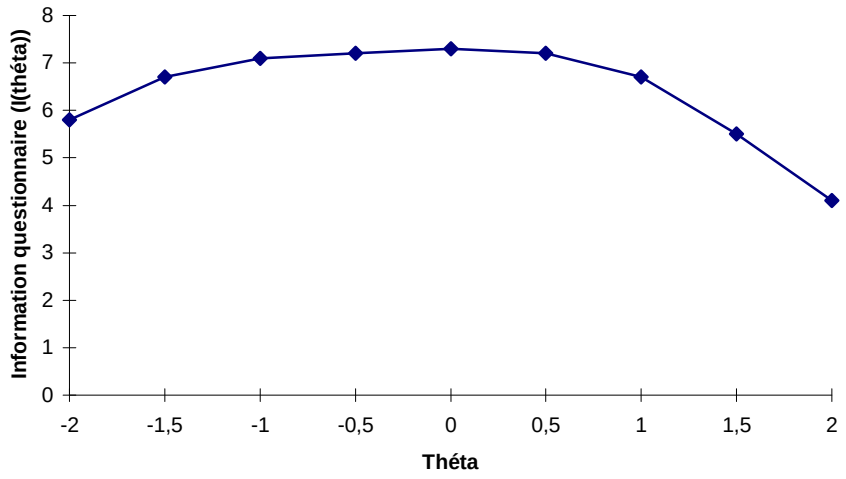
Nous avons construit les courbes d'information des douze items du questionnaire qui nous sert d'exemple. Les courbes caractérisant les items 1, 3, 5, 6, 10 et 12, absentes du graphique ci-dessous afin d'en faciliter la lecture, sont toutes situées entre celles des items 11 et 4 (graphique 3) .



Graphique 3 - Fonctions d'information de certains des items du questionnaire « Orientation de but vers la tâche en français »

Les items 7 et 9 sont ceux qui apportent le plus d'information à l'ensemble du questionnaire, essentiellement aux niveaux d'aptitude moyens (courbes en cloche). On remarque que ce sont les items qui ont les pouvoirs discriminants les plus élevés du questionnaire (voir tableau 4). Pour la majorité des items la quantité d'information apportée diffère peu d'un niveau d'aptitude à un autre. Ces items, par rapport aux items 7 et 9, sont caractérisés par des paramètres de discrimination dont les valeurs sont modérées. Ce sont des items qui vont fournir une moins grande quantité d'information que les items 7 et 9, mais sur une plus large étendue du continuum d'aptitudes. Comme attendu, l'item 2 n'apporte quasiment aucune information au questionnaire.

Les courbes d'information peuvent s'additionner pour estimer la quantité totale d'information apportée par une échelle. Quand les informations d'item sont additionnées sur toute une échelle, la courbe résultante est appelée Courbe d'Information du Test. Cette dernière rend compte de la quantité d'information fournie par un test pour des sujets ayant des niveaux d'aptitude différents (graphique 4) :



Graphique 4 - Courbe d'information du questionnaire « Orientation de but vers la tâche »

Dans l'exemple présenté ci-dessus (graphique 4), la courbe d'information du questionnaire étudié révèle que ce dernier est surtout informatif à des niveaux moyens d'aptitude. C'est à des niveaux élevés d'aptitude qu'il apporte le moins d'information.

L'information d'un test présente une relation intéressante avec l'erreur-standard de l'estimation θ. Spécifiquement, l'inverse de la racine carrée de l'information du test est égale à l'erreur-standard conditionnelle de la métrique θ (ES | θ) (4):

$$ES \mid \theta = 1 / (\text{information} \mid \theta)^{½} \qquad (4)$$

Le calcul de ces erreurs-standard est utile car il nous apporte des indications proches de celles fournies par les indices de fidélité dans le cadre de la théorie classique des tests. Par exemple, si une mesure a une valeur d'information de test égale à 16 à $\theta = 2$, alors les scores du sujet à ce niveau d'aptitude ont une erreur-standard de .25, reflétant une grande précision. De même, si l'échelle fournit une information égale à 4 à $\theta = -2$, alors l'erreur-standard pour les sujets à ce niveau d'aptitude est de .50 indiquant une diminution dans la confiance à accorder dans les scores obtenus au test autour de cette valeur. Dans notre exemple, ces erreurs-standard oscillent entre .37 et .49, les valeurs les plus fortes étant situées aux deux extrémités du continuum d'aptitude (tableau 5) :

θ	-2	-1,5	-1	-0,5	0	0,5	1	1,5	2
ES(θ)	0,42	0,39	0,37	0,37	0,37	0,37	0,39	0,42	0,49

Tableau 5 - Erreurs-standard associées à chaque estimation θ (paramètre « aptitude »)

En conclusion à l'ensemble des analyses qui viennent d'être effectuées, on peut dire qu'il existe une grande convergence d'indicateurs visant à dénoncer le comportement déviant de l'item 2 au sein du questionnaire. Si on se réfère à la formulation de cet item (voir tableau 2), on constate que c'est le seul à faire appel à autrui dans sa formulation. Il est probable donc que cet item de par cette particularité relève d'une autre dimension, d'où cette accumulation d'indices allant dans le sens d'un fonctionnement manifestement différent par rapport aux autres items du questionnaire. La solution généralement préconisée dans ce cas est l'élimination de l'item incriminé.

CONCLUSION

On s'est intéressé ici exclusivement aux paramètres d'items calculés par le Modèle de Réponses Graduées de Samejima. Il ne faut pas oublier cependant que celui-ci permet le calcul de paramètres individuels présentant également un intérêt certain. L'utilisation, lors d'analyses ultérieures, de ces scores théoriques (paramètres « aptitude ») plutôt que de scores bruts présente l'avantage de pouvoir différencier le niveau d'aptitude de deux sujets ayant obtenu deux mêmes scores bruts à l'aide de combinaisons différentes de réponses (des patterns de réponses différents). Ce rééchelonnement des scores bruts totaux permet généralement une description plus juste des données. Le modèle fournit en effet autant de paramètres « aptitude » qu'il y a de patterns différents de réponses au sein d'un questionnaire et permet donc une représentation affinée des données. C'est cette caractéristique qui nous a fait d'ailleurs essentiellement préférer ce modèle aux extensions du modèle de Rasch pour items polytomiques. En effet, ces derniers procèdent un peu différemment en se basant lors du rééchelonnement sur les scores totaux sans tenir compte du fait que plusieurs patterns de réponses différents peuvent résulter en un même score total.

Les MRI présentent, entre autres, les avantages suivants : 1) ils offrent la possibilité de construire des instruments de mesure comportant très peu d'items, 2) ils permettent un examen minutieux du comportement de chaque item, ce qui se révèle particulièrement important quand le nombre d'items étudiés est faible (un item déviant dans ce cas peut fausser considérablement les analyses), 3) ils nous aident à mieux comprendre ce que l'échelle considérée mesure (Steinberg & Thissen, 1995). Cependant, malgré les qualités psychométriques très attractives de ces modèles, certains auteurs (Hox & Mellenbergh, 1990) mettent en garde contre une application qui tendrait à devenir routinière de ceux-ci. En effet, il ne faut pas oublier que ces modèles ont pour point de départ une certaine théorie des

processus de réponses (Lord, 1980). Beaucoup d'entre eux ont été développés à l'origine pour l'étude des tests d'intelligence et dans ce cadre, un modèle qui déclare que la probabilité de donner une réponse correcte devient plus faible quand les questions deviennent plus difficiles a du sens. Par contre, quand les items ne renvoient pas à des tests d'intelligence mais à des questions sur les attitudes, les valeurs, ..., la pertinence de l'utilisation de ces modèles doit toujours préalablement être questionnée.

BIBLIOGRAPHIE

Andersen, E.B. (1997). The rating scale model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 67-84). New York: Springer.

Andrich, D. (1995). Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measurement, 19*, 1, 101-119.

De Ayala, R.J. (1993). An introduction to polytomous item response theory models. *Measurement and Evaluation in Counseling and Development, 25*, 172-189.

Flannery, W.M.P., Reise, S.P. & Widaman, K.F. (1995). An item response theory analysis of the general and academic scales of the self-description questionnaire II. *Journal of Research in Personality, 29*, 168-188.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamendals of item response theory.* Newbury Park: Sage Publications Inc.

Hox, J.J. & Mellenbergh, G.J. (1990). The interplay of substantive and auxiliary theory. In J.J. Hox & J. De Jong-Gierveld (Eds.), *Operationalization and research strategy* (pp 123-136). Amsterdam: Swets & Zeitlinger.

Koch, W.R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7*, 15-32.

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

Masters, G.N. & Wright, B.D. (1997). The partial credit model. In W.J. van der Linden & R.K. Hambleton (eds.), *Handbook of modern Item Response Theory* (pp. 101-121). New York: Springer.

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

Samejima, F. (1997). Graded response model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern Item Response Theory* (pp. 85-100). New York: Springer.

Seegers, G. & Boekaerts, M. (1993). Task motivation and mathematics achievement in actual task situations. *Learning and Instruction, 3*, 133-150.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350-353.

Steinberg, L. & Thissen, D. (1995). Item response theory in personality research. In P.E. Shrout & S.T. Fiske (Eds.), *Personality research, methods and theory: festschrift honoring Donald W. Fiske* (pp. 161-181). Hillsdale, N.J.: Erlbaum.

Thissen, D. (1991). *MULTILOG User's Guide : Multiple, categorical item analysis and test scoring using item response theory* (Version 6.0). Chicago : Scientific Software, INC.

Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.

Thissen, D. & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*, 385-395.

Thissen, D. & Steinberg, L. (1997). A response model for multiple-choice items. In W.J. van der Linden & R.K. Hambleton (eds), *Handbook of modern Item Response Theory* (pp. 61-65). New York: Springer.

Van der Linden, W.J. & Hambleton, R.K. (1997). Item response theory: brief history, common models and extensions. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern Item Response Theory* (pp. 1-28). New York: Springer.

Weiss, D.J. & Yoes, M.E. (1991). Item Response Theory. In R.K. Hambleton & J.N. Zaa (Eds.), *Advances in educational and psychological testing* (pp. 69-95). Kluwer Academic Publishers.