Mesure et évaluation en éducation

L'impact de la formulation des items dans les questionnaires d'enquête

Une étude avec le modèle de Rasch pour les données polytomiques

Jean-Guy Blais and Julie Grondin

Volume 33, Number 2, 2010

URI: https://id.erudit.org/iderudit/1024897ar
DOI: https://doi.org/10.7202/1024897ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

Blais, J.-G. & Grondin, J. (2010). L'impact de la formulation des items dans les questionnaires d'enquête : une étude avec le modèle de Rasch pour les données polytomiques. *Mesure et évaluation en éducation*, *33*(2), 95–126. https://doi.org/10.7202/1024897ar

Article abstract

Self-administered survey questionnaires are a primary source of data collection in education and in the social sciences in general. It is often difficult to create a question which will be interpreted in the same way by all respondents and then meet the researcher's goals. A change of wording can influence the meaning of an item, but it is also possible that there would be no influence at all; it all depends on the changes made. Besides, since the first studies related to items' wording in survey questionnaires, new measurement models were developed. This is the case notably for measurement models from the Rasch family. The goal of this paper is to illustrate how the developments surrounding a measurement model from the Rasch family, namely Andrich's model for polytomous data, can be put at contribution to study the relative impact of items' wording on the responses obtained in survey questionnaires.

L'impact de la formulation des items dans les questionnaires d'enquête : une étude avec le modèle de Rasch pour les données polytomiques

Jean-Guy Blais
Université de Montréal

Julie Grondin
UQAR, campus de Lévis

Mots clés : Questionnaire, formulation des items, modèle de Rasch, données polytomiques

Les enquêtes par questionnaire auto-administré sont un des principaux instruments de cueillette de données utilisés par les chercheurs en éducation et en sciences sociales en général. Or, il peut être difficile de créer une question qui sera interprétée de la même façon par tous les répondants et qui répondra ainsi aux objectifs des chercheurs. Un changement de mots peut avoir un impact sur la signification d'un item, mais il est aussi possible qu'il n'y ait pas d'impact ; tout dépend du changement apporté. Par ailleurs, depuis les premières études portant sur la formulation des items dans les enquêtes par questionnaire, de nouveaux modèles de mesure ont été élaborés. C'est le cas, notamment, de la famille des modèles de Rasch. L'objectif de cet article est d'illustrer comment les développements entourant un modèle de mesure de la famille des modèles de Rasch, notamment le modèle polytomique de Andrich, peuvent être mis à contribution pour étudier la portée relative de la formulation des items sur les réponses obtenues à l'aide de questionnaires d'enquête.

Key words: Questionnaires, items' wording, Rasch model, polytomous data

Self-administered survey questionnaires are a primary source of data collection in education and in the social sciences in general. It is often difficult to create a question which will be interpreted in the same way by all respondents and then meet the researcher's goals. A change of wording can influence the meaning of an item, but it is also possible that there would be no influence at all; it all depends on the changes made. Besides, since the first studies related to items' wording in survey questionnaires, new measurement models were developed. This is the case notably for measurement models from the Rasch family. The goal of this paper is to illustrate how the developments surrounding a measurement model from the

Rasch family, namely Andrich's model for polytomous data, can be put at contribution to study the relative impact of items' wording on the responses obtained in survey questionnaires.

Palavras-chave: Questionário, formulação de itens, modelo de Rasch, dados politómicos

Os inquéritos por questionário auto-administrados são um dos principais instrumentos de recolha de dados utilizados pelos investigadores em educação e nas ciências sociais em geral. Ora, é difícil conceber uma questão que seja interpretada da mesma maneira por todos os respondentes e que, desse modo, responda aos objectivos dos investigadores. Uma mudança de palavras pode ter um impacto sobre o significado de um item, mas também pode acontecer que não tenha impacto nenhum, tudo dependendo da mudança realizada. De resto, depois dos primeiros estudos sobre a formulação de itens nos inquéritos por questionário, foram elaborados novos modelos de medida, como é o caso da família de modelos de Rasch. O objectivo deste artigo é ilustrar como os desenvolvimentos em torno de um modelo de medida da família de modelos de Rasch, designadamente o modelo politómico de Andrich, pode dar um contributo no estudo do impacto relativo da formulação de itens sobre as respostas obtidas com a ajuda de inquéritos por questionário.

Note des auteurs – La recherche dont certains des résultats sont présentés ici a été subventionnée par le Conseil de recherche en sciences humaines du Canada (CRSH) de 2002 à 2005 et de 2005 à 2008. Toute correspondance peut être adressée comme suit : Jean-Guy Blais, professeur titulaire, Département d'administration et fondements de l'éducation, Faculté des sciences de l'éducation, Université de Montréal, C.P. 6128 Succursale Centre-ville, Montréal (Québec) H3C 3J7, ou Julie Grondin, professionnelle de recherche, Département des sciences de l'éducation, UQAR, campus de Lévis, 1595, boulevard Alphonse-Desjardins, Lévis (Québec) G6V 0A6, ou par courriel aux adresses suivantes : [jean-guy.blais@umontreal.ca] ou [julie_grondin@uqar.qc.ca].

Introduction

Les questionnaires d'enquête auto-administrés constituent un des principaux instruments de cueillette de données utilisés par les chercheurs en éducation et en sciences sociales en général. Le plus souvent, les chercheurs souhaitent mieux comprendre les motivations ou les processus cognitifs qui sont à la base des attitudes ou des opinions des répondants (Abelson, 1992; Krosnick & Abelson, 1992; Schwarz, 1999). Or, de nombreux facteurs peuvent nuire à la compréhension des items par les répondants et, par conséquent, avoir un impact sur la qualité des données recueillies.

Selon Tourangeau et Rasinski (1988), tous les éléments précédant l'item auquel un sujet interrogé tente de répondre forment un cadre interprétatif qui lui permet de déduire la perspective qu'il doit donner à sa réponse. Ainsi, les aspects d'un questionnaire tels que son format, le texte d'introduction, le contexte, la formulation des items, le format des réponses, le nombre et les catégories de réponses proposées deviennent des indices qui permettent au répondant de déterminer l'intention du chercheur et, par suite, constituent des sources d'influence sur l'interprétation que font les répondants de ce qui leur est demandé (Clark & Schober, 1992; Groves, Fultz, & Martin, 1992; Schaeffer & Presser, 2003; Schwarz, 1999). Parmi ces facteurs, l'influence de la formulation des items demeure un des éléments au sujet duquel les connaissances sont les plus limitées. Certains chercheurs ont pu montrer que le choix des mots utilisés pour la formulation des items exerçait effectivement une influence sur les réponses des participants (Blais, Grondin, Loye, & Raîche, 2007; Kahneman & Tversky, 1982; Ng, Pipe, Beath, & Holton, 1999). Cependant, l'ampleur de cette influence demeure difficile à évaluer. Certains changements de mots, mêmes mineurs, peuvent avoir des impacts majeurs sur les résultats d'une enquête, alors que des changements majeurs peuvent n'avoir aucun impact. Il apparaît donc important de poursuivre les recherches sur les caractéristiques d'une enquête par questionnaire, telle la formulation des items, qui contribuent à la contamination des données recueillies.

Pour ce faire, il semble pertinent d'utiliser un modèle de mesure qui permet une étude plus approfondie des caractéristiques métrologiques des items. Depuis les premières études portant sur la formulation des items dans les enquêtes par questionnaire, de nouveaux modèles de mesure ont été élaborés. C'est le cas notamment de la famille des modèles de Rasch. Certains de ces modèles, comme le modèle polytomique pour les catégories de réponses ordonnées (ou modèle *Rating Scale*) de Andrich (1978), ont été proposés de façon à permettre aux chercheurs de modéliser les données obtenues à l'aide d'échelles de réponses de type Likert et qui sont souvent utilisées dans les enquêtes par questionnaire. Ce modèle pallierait aux principales critiques généralement formulées à l'égard de la théorie classique des tests : lorsqu'il est adéquat, il permettrait aux chercheurs d'obtenir des résultats invariants et des scores avec des unités de mesures égales.

L'objectif de cet article est donc d'illustrer comment les développements entourant un modèle de mesure de la famille des modèles de Rasch, notamment le modèle polytomique de Andrich (1978), peuvent être mis à contribution pour étudier la portée relative de la formulation des items sur les réponses obtenues à l'aide d'enquêtes par questionnaire.

La formulation des items

Les objets d'étude des sciences sociales sont complexes. Il est souvent difficile de créer une question qui sera interprétée de la même façon par tous les répondants et qui respecte leurs sentiments. Un changement de mots en apparence simple peut parfois provoquer des différences substantielles sur la signification d'un item. Pour illustrer ce propos, Schuman et Kalton (1985) présentent deux items portant sur le contrôle des armes à feu. Dans un premier cas, l'item est formulé de façon à obliger les individus d'une population à obtenir un permis auprès d'un service policier avant l'achat d'une arme. Dans le deuxième cas, l'item est formulé de façon à interdire l'achat d'une arme par les individus d'une population, sauf pour les autorités policières ou militaires. Les deux formulations portent sur un même concept, mais les termes clés utilisés sont substantiellement différents et risquent d'engendrer des réponses différentes de la part des répondants. En effet, une vaste majorité de répondants sera généralement en faveur de *permettre*, moyennant un contrôle des armes à feu, alors que, le plus souvent, elle s'opposera au fait d'*interdire* (Rugg, 1941; Smith, 1980).

Des changements de mots moins importants, comme ne changer qu'un seul mot ou une seule lettre, sont également susceptibles d'engendrer des réponses différentes de la part des répondants. Par exemple, si le mot *achat* était remplacé par le mot *utilisation* dans l'exemple précédent, il est facile d'imaginer qu'une vaste majorité de la population serait en faveur d'*interdire* l'utilisation des armes à feu. Cependant, tous les changements de mots ne provoquent pas nécessairement des différences dans les données recueillies. C'est d'ailleurs là tout le problème : tous les changements de mots n'affectent pas les réponses, mais dans des contextes précis, certains changements de mots peuvent avoir un impact non négligeable. Il est donc difficile de prédire à l'avance quels changements de mots auront un impact sur les résultats et lesquels n'en auront pas. Bien qu'il soit possible d'imaginer que l'impact d'un changement de mots puisse être presque illimité (c'est-à-dire provoquer des différences de presque 100 % dans la distribution des réponses obtenues avec des modifications appropriées), la plupart des changements provoquent des différences de moins de 20 %, et souvent, de moins de 10 %. Les différences trouvées entre *permettre* et *interdire* seraient les plus grandes que la recherche ait pu trouver (Schuman & Kalton, 1985).

Toujours selon Schuman et Kalton, il convient donc de s'interroger sur l'effet d'une telle différence pour la recherche. Si un certain pourcentage de la population endosse une attitude, il est très improbable qu'un changement de mots non substantiel provoque une différence de plus de 20 % dans la distribution des réponses obtenues. Dans un tel cas, il est vrai que la précision des résultats et la position relative des items peuvent en être affectées. En revanche, si une majorité endossait l'attitude en question, il apparaît invraisemblable que cette majorité en vienne à rejeter l'attitude en question. On pourrait donc compter sur le fait que, dans la plupart des cas, les variations seraient suffisamment minces pour ne pas entraîner de distorsions sérieuses dans les résultats et que les items pourraient être considérés équivalents (Blais, 1992).

Le modèle de mesure

Les enquêtes par questionnaire qui utilisent des échelles de réponses en catégories ordonnées pour quantifier une attitude sont très nombreuses puisque ce type d'échelle est très facile à utiliser. Plusieurs critiques leur sont cependant associées. Tout d'abord, elles n'offrent aucune possibilité de comparaisons directes ni proportionnelles entre les données recueillies (Borg, 1982). Il est possible de faire une certaine estimation du niveau *fort* ou *faible* de l'attitude

des gens, mais il n'est pas possible d'avoir une estimation plus précise, relative à une unité commune, et qui permette d'évaluer de « combien plus fort » ou « de combien plus faible » est cette attitude.

Ensuite, un item peut être plus facile ou plus difficile à endosser selon le groupe de participants interrogés (Hambleton, Swaminathan, & Rogers, 1991). Ainsi, l'attitude des répondants, recueillie à l'aide de ces échelles de réponses et estimée à l'aide du modèle de la théorie classique des tests (TCT), dépend du groupe de sujets interrogés (Blais & Ajar, 1992). Avec ce modèle, les comparaisons entre l'attitude de deux groupes de répondants devraient donc se limiter à des situations où les conditions entourant le processus de mesure sont les mêmes et où les groupes de candidats possèdent des caractéristiques similaires.

Enfin, les catégories de réponse des échelles utilisées sont généralement ordonnées, mais les intervalles entre chacune des catégories sont inégaux (Blais & Grondin, soumis ; Bradburn & Sudman, 1979 ; Cools, Hofmans, & Theuns, 2006 ; Rohrmann, 2003 ; Tourangeau, Rips, & Rasinski, 2000). En conséquence, l'estimation de l'attitude des gens obtenue avec le modèle de la TCT ne permet pas de résumer de façon adéquate les données recueillies. Pour qu'une mesure soit utile, elle doit permettre aux chercheurs de faire des prédictions (Dawes & Smith, 1985; Thissen & Wainer, 2001), d'en estimer la précision et de détecter ainsi qu'évaluer les divergences entre les données recueillies et les mesures modélisées (Wright & Mok, 2004). C'est précisément ce que permettent de faire les modèles de mesure de la famille des modèles de Rasch.

Un modèle de Rasch polytomique pour des catégories ordonnées

Les modèles de la famille de Rasch font partie d'une catégorie de modèles de mesure abondamment étudiés ces 30 dernières années, mais somme toute encore peu utilisés pour des applications dans le domaine de la recherche. Parmi les avantages importants que procurent les modèles de Rasch, l'invariance des paramètres est celui qui est le plus recherché. Lorsque les données s'ajustent adéquatement au modèle, les estimations des valeurs de la variable latente du modèle (par exemple l'attitude des gens) possèdent la propriété d'invariance, c'est-à-dire que ce sont des estimations qui sont indépendantes du groupe de personnes ciblées par l'opération de mesure ou du groupe d'items inclus dans l'instrument. Aussi, la modélisation de Rasch repose sur un processus stochastique qui permet de transformer les données brutes, *i.e.* les réponses sur l'échelle de consignation, en données sur une

échelle dont les intervalles entre les unités de mesure sont égaux (Bond & Fox, 2001). Enfin, le modèle polytomique de Andrich (1978) (aussi appelé le *Rating Scale Model*), conçu pour les échelles de réponses en catégories ordonnées, est un modèle unidimensionnel qui s'applique lorsque l'hypothèse de la présence d'une seule variable latente est plausible, lorsqu'il y a plus de deux catégories de réponses et lorsque l'intervalle entre chaque catégorie de réponses de l'échelle demeure le même pour tous les items du questionnaire. Ce modèle transforme l'échelle en une succession de situations dichotomiques. Il se sert du point où la probabilité d'opter pour la prochaine catégorie de réponse est égale à celle de conserver la précédente de telle sorte que ces intervalles peuvent être interprétés comme le succès ou l'échec de passer à la catégorie suivante. Une représentation possible du modèle est :

$$P_{ni} = \frac{e^{(Bn - Di - Fx)}}{1 + e^{(Bn - Di - Fx)}}$$

Lorsqu'on prend le logarithme naturel du rapport des chances, on obtient :

$$Ln\,[P_{nix} / P_{ni(x-1)}] = B_n - D_i - F_x$$

où P_{nix} représente la probabilité que la personne n avec une attitude B_n adhère à la catégorie x (où $x = 0$ à m-1, pour m catégories de réponses offertes) d'un item i qui se situe à la position D_i du continuum de l'endossement. Le paramètre F_x correspond au seuil entre les catégories x-1 et x ou, plus précisément, au point où la probabilité d'opter pour l'une ou l'autre des catégories est égale. F_x peut également être interprété comme l'écart entre la catégorie x-1 et la catégorie x. Enfin, $P_{ni(x-1)}$ représente la probabilité que la personne n avec une attitude B_n endosse la catégorie x-1. Dans une enquête par questionnaire auto-administré sur les attitudes, B_n correspond au degré d'endossement du répondant n par rapport à l'attitude évaluée par l'ensemble des items proposés et D_i, à la difficulté qu'éprouvent les personnes à endosser l'item i. Un item dont le niveau de difficulté est élevé est donc un item difficile à endosser et une personne dont le degré d'endossement est élevé est une personne généralement favorable à l'attitude visée par les items proposés.

L'ajustement des données au modèle

De façon générale, les statistiques d'ajustement permettent aux chercheurs de déterminer si le modèle de mesure choisi est effectivement approprié pour représenter les données. Cependant, le chercheur qui travaille avec la famille des modèles de Rasch doit faire une démonstration différente. En effet, les modèles de la famille de Rasch sont des modèles théoriques qui, lorsque les

conditions de base sont respectées, permettent d'obtenir des estimations de paramètres utiles du point de vue de la mesure (Bond & Fox, 2001). En conséquence, ce sont les données qui doivent s'ajuster au modèle et non le contraire. Les indices d'ajustement des modèles de Rasch permettent ainsi aux chercheurs de distinguer les données qui ne s'ajustent pas bien au modèle et, par conséquent, qui ne satisfont pas aux conditions de base du modèle. Par exemple, les items qui ne respectent pas l'hypothèse d'unidimensionnalité du modèle, c'est-à-dire qui divergent du modèle théorique attendu, ne s'ajusteront pas bien au modèle (Bond & Fox, 2001).

Différents indices statistiques sont proposés pour la détection de sources de discordances entre les données et le modèle. Ces indices peuvent être calculés de deux façons (Smith, 2004). D'une part, ils peuvent être calculés à partir de la matrice des résidus standardisés mis au carré. Cette première méthode de calcul correspond à la version non pondérée (*unweighted*) des statistiques d'ajustement ou indice *outfit (outlier sensitive mean square residual goodness of fit statistic)*. Cet indice permet de détecter les données aberrantes (qui s'éloignent du patron de réponse attendu). Dans la deuxième version de l'indice, chacun des éléments de la matrice standardisée mis au carré est divisé par la fonction d'information. L'effet de cette pondération est ensuite annulé en divisant la somme obtenue sur chacune des colonnes par la somme des poids utilisés. Cette méthode de calcul correspond à la version pondérée (*weighted*) ou indice *infit (information weighted mean square residual goodness of fit statistic)*. Dans cette version de l'indice, plus de poids est accordé aux personnes dont le niveau d'endossement envers une attitude est proche du niveau d'endossement de l'item (et vice versa). Cet indice permet donc d'étudier les patrons de réponse inattendus près du niveau d'endossement de l'item (ou de la personne). En conséquence, les problèmes d'ajustement détectés par la statistique *infit* sont généralement plus difficiles à diagnostiquer et à corriger. Ils présentent donc un risque plus grand pour la mesure. Par contre, la statistique *outfit* permet d'obtenir une puissance de test plus élevée que la statistique *infit*; son taux d'erreur de type I est également plus stable que celui de l'*infit*; et il est moins sensible à différentes tailles d'échantillon (Smith, 2004). L'indice *infit* serait effectivement influencé par la taille de l'échantillon, la longueur de l'outil d'évaluation, de même que par le nombre d'options offertes dans les catégories de réponses (Curtis, 2003, tel que cité par Curtis et Boman, 2004). Chacun de ces indices ayant été formulé dans le but de détecter un type particulier de divergence entre les données et le modèle (Linacre, 1996), il convient donc de les utiliser de façon complémentaire.

Les indices *infit* et *outfit* peuvent être interprétés comme des statistiques ayant une distribution khi-carré. Ainsi, pour deux degrés de liberté différents, la valeur de la région critique associée sera également différente. Il devient alors impossible de déterminer une valeur unique comme point de référence afin de juger de la qualité de l'ajustement entre les données et le modèle (Smith, 2004). Pour faciliter l'interprétation de ces indices, deux transformations sont principalement utilisées. La première consiste à diviser ces indices par leur nombre de degrés de liberté. Les statistiques sont ainsi transformées en carré moyen (*mean square*). La seconde consiste à transformer le carré moyen en racine cubique. Cette transformation permet de convertir le carré moyen en une statistique qui s'apparente à la statistique t de Student. Cette statistique est communément appelée l'indice d'ajustement standardisé (*standardized fit index*). La distribution de l'indice d'ajustement standardisé posséderait des propriétés qui le rendent plus stable que le carré moyen par rapport à des tailles d'échantillon différentes (Smith, Schumacker, & Bush, 1998; Wang & Chen, 2005). La version standardisée de ces indices permettrait également de détecter davantage de problèmes d'ajustement que le carré moyen (Smith & Suh, 2003). C'est donc la version standardisée qui a été retenue pour cette étude.

Enfin, il n'y a pas de réponse unique sur la grandeur que doit avoir la valeur d'un indice statistique pour être considérée comme ne s'ajustant pas bien au modèle. De façon générale, une valeur d'indice *infit* ou *outfit* standardisés se situant à l'extérieur de l'intervalle [-2 ; 2] est considérée comme révélatrice d'un problème d'ajustement. Cependant, Li et Olejnik (1997) ont trouvé que la distribution de ces indices déviait de celle d'une distribution normale et que, en conséquence, ces indices auraient besoin d'un peu plus de latitude que l'intervalle habituel [-2 ; 2]. Considérant la taille des groupes obtenus dans cette étude et les objectifs qui sont de montrer l'utilité des outils de détection de la qualité de l'ajustement liés aux modèles de Rasch pour ce type d'analyse, un intervalle de [-2,5 ; 2,5] a donc été retenu (Lawton, Bhakta, Chamberlain, & Tennant, 2004).

L'instrument de cueillette des données

En 1999, le Centre de formation initiale des maîtres (CFIM) de l'Université de Montréal a mis au point un questionnaire d'enquête auto-administré afin d'obtenir des données pour évaluer ses programmes de premier cycle universitaire en formation des maîtres. Le questionnaire a été distribué pour la première fois au printemps de l'an 2000 et une opération de récolte de données a eu lieu chaque année depuis, mais avec deux formes d'un même questionnaire. La dernière récolte de données effectuée dans le cadre de ce projet de recherche a eu lieu en 2007.

Le questionnaire original de 2000 était composé de huit sections et il a été modifié avec les années de sorte que la dernière version utilisée dans le cadre de cette recherche ne comportait plus que quatre sections (perception générale de la formation, préparation à l'enseignement, les stages et divers renseignements d'ordre démographique). Seule la section concernant la préparation à l'enseignement a été retenue comme objet d'étude pour la recherche menée. Dans cette section du questionnaire, les étudiants doivent répondre à 20 items introduits par la phrase « Je considère que mon programme d'études m'a permis de développer des compétences pour… ». L'échelle de réponses fournie aux étudiants pour les années qui font l'objet de cette présentation, *i.e.* 2004 et 2005, est une échelle de type Likert en six points. L'échelle est strictement positive et les catégories de réponses sont identifiées par des valeurs numériques. Seules la première et la dernière catégorie possèdent une étiquette. Ainsi, les catégories de réponses vont de 1 (*Tout à fait en désaccord*) à 6 (*Tout à fait en accord*).

Au printemps de l'année 2004, deux versions du questionnaire ont été distribuées (comme à chaque année de l'étude transversale 2000-2007). La version A du questionnaire est la version de référence et la version B, celle qui a subi des changements. Le tableau 1 montre que, dans cette dernière, la formulation a été modifiée pour 11 des 20 items. Les modifications à la formulation ont été apportées en fonction de deux critères : l'ajustement statistique des items au modèle observé lors des années précédentes, et une formulation plus précise de certains items. Ainsi, la formulation de cinq items a été changée parce que l'étude de l'ajustement de ces items révélait un problème potentiel et des changements de mots ont été appliqués à cinq autres items de façon à en rendre l'objet plus précis. Un des items a cependant été reformulé de façon à ce que le sujet soit plus général (item 7).

Tableau 1
Formulation des items pour les versions A et B du questionnaire 2004

2004 A	2004 B
1. Identifier les contenus difficiles **à faire apprendre aux** élèves	1. Identifier les contenus **plus** difficiles **pour les** élèves
2. Répondre aux questions **des parents lors de la présentation du bulletin**	2. Répondre aux questions **lors des rencontres avec les parents**
3. **Construire** des outils pour l'évaluation sommative **(contrôles, examens, etc.)**	3. **Élaborer** des outils pour l'évaluation sommative
4. Maîtriser les contenus que j'enseignerai **en conformité avec les programmes du ministère de l'Éducation**	4. Maîtriser les contenus que j'enseignerai
5. Corriger **la langue écrite** des élèves	5. Corriger **les productions écrites** des élèves
6. Corriger la langue orale des élèves	6. Corriger la langue orale des élèves
7. **Intervenir individuellement auprès des élèves à risque d'échouer**	7. **Planifier une intervention auprès d'un élève en difficulté**
8. Planifier le déroulement d'activités d'apprentissage	8. Planifier le déroulement d'activités d'apprentissage
9. Adapter mes activités d'enseignement aux caractéristiques des élèves	9. Adapter mes activités d'enseignement aux caractéristiques des élèves
10. Établir les règles de fonctionnement de la classe	10. Établir les règles de fonctionnement de la classe
11. Motiver les élèves **à s'engager** dans leur apprentissage	11. Motiver les élèves dans leurs apprentissages
12. Respecter les différences ethniques ou culturelles des élèves	12. Respecter les différences ethniques ou culturelles des élèves
13. **Identifier** les points forts et les points faibles des élèves	13. **Reconnaître** les points forts et les points faibles des élèves
14. Collaborer avec les autres enseignantes et enseignants	14. Collaborer avec les autres enseignantes et enseignants
15. **Construire** des outils pour l'évaluation formative **(exercices, devoirs, etc.)**	15. **Élaborer** des outils pour l'évaluation formative
16. Aider les élèves à développer leurs méthodes de travail	16. Aider les élèves à développer leurs méthodes de travail
17. Sanctionner les problèmes de discipline chez les élèves	17. Sanctionner les problèmes de discipline chez les élèves
18. Sensibiliser les élèves aux situations de discrimination qui existent entre eux	18 Sensibiliser les élèves aux situations de discrimination qui existent entre eux
19. **Orienter les** élèves **vers les** services d'aide appropriés	19. **Suggérer aux** élèves **des** services d'aide appropriés
20. **Discuter avec les** parents **des** difficultés de leur enfant	20. **Expliquer aux** parents **quelles sont les** difficultés de leur enfant

La qualité de l'ajustement entre les données et le modèle

La cueillette de données de 2004 a permis de récolter les réponses de 125 étudiants et étudiantes du B.EPEP. Parmi ces 125 personnes, 63 ont répondu à la version A du questionnaire et 62 ont répondu à la version B. Pour sa part, la cueillette de données de 2005 a permis de recueillir les réponses de 113 étudiants et étudiantes. Parmi ces 113 personnes, 72 ont répondu à la version A du questionnaire et 41 ont répondu à la version B.

Tejada, Gómez, García et Meléndez (2002) ont trouvé qu'amorcer l'analyse de la qualité de l'ajustement par l'étude de l'ajustement des personnes permettait de conserver plus d'items (et vice versa). C'est donc la stratégie qui a été utilisée. L'analyse de l'ajustement des personnes a révélé que 11 personnes ayant répondu à la version A du questionnaire 2004 présentent des patrons de réponses produisant des indices d'ajustement élevés, de même que neuf personnes ayant répondu à la version B. De plus, la corrélation entre la mesure estimée par le modèle et les données recueillies est négative pour une des personnes ayant répondu à la version B du questionnaire. La corrélation produite par Winsteps est une forme de corrélation point-bisériale. Ainsi, une valeur négative révèle généralement la présence de données aberrantes dans les réponses d'un individu ou la présence d'un item dont les réponses sont inversées par rapport à l'échelle proposée. Pour l'année 2005, c'est le patron de réponses de 12 personnes ayant répondu à la version A du questionnaire qui ne s'ajuste pas bien au modèle, de même que le patron de réponses de huit personnes ayant répondu à la version B. De plus, la valeur de la corrélation calculée par Winsteps est négative pour un des participants ayant répondu à la version A du questionnaire.

Ici, les échantillons utilisés sont petits. Cependant, lorsque les échantillons utilisés sont très grands, il est impensable de vérifier chacun des patrons de réponses afin de déterminer si ceux-ci contribuent de façon utile ou non à l'estimation des paramètres par le modèle (Curtis & Boman, 2004). Il devrait alors être possible de se fier uniquement aux indices d'ajustement afin de déterminer si un patron de réponse contrevient à la mesure. Étant donné les visées méthodologiques de cette étude, c'est la façon de procéder qui a été utilisée. Cependant, à titre exploratoire, les patrons de réponse des sujets qui ne s'ajustent pas au modèle ont tout de même été examinés afin de tenter de déterminer pourquoi ils ne s'ajustent pas au modèle et s'ils nuisent à la mesure (Smith, 1996). Puisque rien ne nous permettait de découvrir la source du problème ou l'action à prendre pour le corriger, nous avons décidé de retirer

ces sujets (Linacre & Wright, 1994). En effet, conserver les sujets qui ne s'ajustent pas au modèle pourrait diminuer la précision de la mesure (Curtis & Boman, 2004).

L'analyse de la qualité de l'ajustement étant un processus itératif, les personnes pour lesquelles le patron de réponses a produit les pires indices d'ajustement standardisés ou une corrélation négative ont été retirées de l'échantillon. Après chacun des retraits, la qualité de l'ajustement était de nouveau étudiée. Ce processus a fait en sorte que certains patrons de réponses qui ne s'ajustaient pas bien au modèle initialement se sont mieux ajustés au modèle par la suite. De même, le processus a permis de faire ressortir de nouveaux patrons de réponses qui ne s'ajustaient plus au modèle. Au total, 12 personnes ont été retirées de l'échantillon pour la version A du questionnaire 2004 et dix personnes pour la version B. Pour l'année 2005, ce sont 11 personnes qui ont été retirées de l'échantillon pour la version A du questionnaire et dix personnes pour la version B. En conséquence, les réponses de 51 personnes ont été conservées dans l'échantillon pour la version A du questionnaire 2004 et celles de 52 personnes pour la version B. Pour le questionnaire 2005, 61 personnes ont été conservées pour la version A du questionnaire et 31 personnes pour la version B.

L'analyse de la qualité de l'ajustement pour les items a ensuite permis de constater, tel qu'on peut l'observer au tableau 3, que dans la version A du questionnaire 2004, les réponses fournies à deux items ne s'ajustent pas bien au modèle (les items 12 et 13). Notons que l'item 13 est un des items dont la formulation a été modifiée. Dans la version B du questionnaire, tous les items s'ajustent bien au modèle. Pour l'année 2005, les patrons de réponses de trois items ne s'ajustent pas bien au modèle (items 1, 3 et 13) pour la version A du questionnaire. Dans la version B, ce sont les patrons des items 6, 8 et 19 qui ne s'ajustent pas bien au modèle.

Tableau 3
Statistiques d'ajustement des items pour les versions A et B des questionnaires 2004 et 2005

2004 A							
Item	**Mesure estimée**	**Erreur type**	**Infit MNSQ**	**Infit ZSTD**	**Outfit MNSQ**	**Outfit ZSTD**	**Corr.**
Items *misfit*							
12	**-1,46**	**0,17**	**1,58**	**2,6**	**1,54**	**2,5**	**0,55**
13	**-0,30**	**0,16**	**0,57**	**-2,6**	**0,55**	**-2,8**	**0,79**
Ensemble des items							
MOY.	**0,00**	**0,17**	**0,99**	**-0,1**	**0,98**	**-0,2**	
É.-T.	**0,92**	**0,01**	**0,25**	**1,3**	**0,25**	**1,3**	

2005 A							
Item	**Mesure estimée**	**Erreur type**	**Infit MNSQ**	**Infit ZSTD**	**Outfit MNSQ**	**Outfit ZSTD**	**Corr.**
Items *misfit*							
1	**0,64**	**0,16**	**0,60**	**-2,6**	**0,60**	**-2,6**	**0,69**
3	**0,86**	**0,16**	**1,60**	**3,0**	**1,59**	**2,9**	**0,56**
6	-0,38	0,16	1,08	0,5	1,10	0,6	0,62
8	-2,01	0,18	1,20	1,1	1,17	1,0	0,55
13	**-0,03**	**0,16**	**0,53**	**-3,2**	**0,53**	**-3,2**	**0,74**
19	0,83	0,16	1,04	0,3	1,01	0,1	0,70
Ensemble des items							
MOY.	**0,00**	**0,16**	**0,99**	**-0,1**	**0,99**	**-0,10**	
É.-T.	**0,73**	**0,01**	**0,28**	**1,6**	**0,28**	**1,6**	

2004 B							
Item	**Mesure estimée**	**Erreur type**	**Infit MNSQ**	**Infit ZSTD**	**Outfit MNSQ**	**Outfit ZSTD**	**Corr.**
Items *misfit*							
12	-1,87	0,18	1,33	1,5	1,35	1,7	

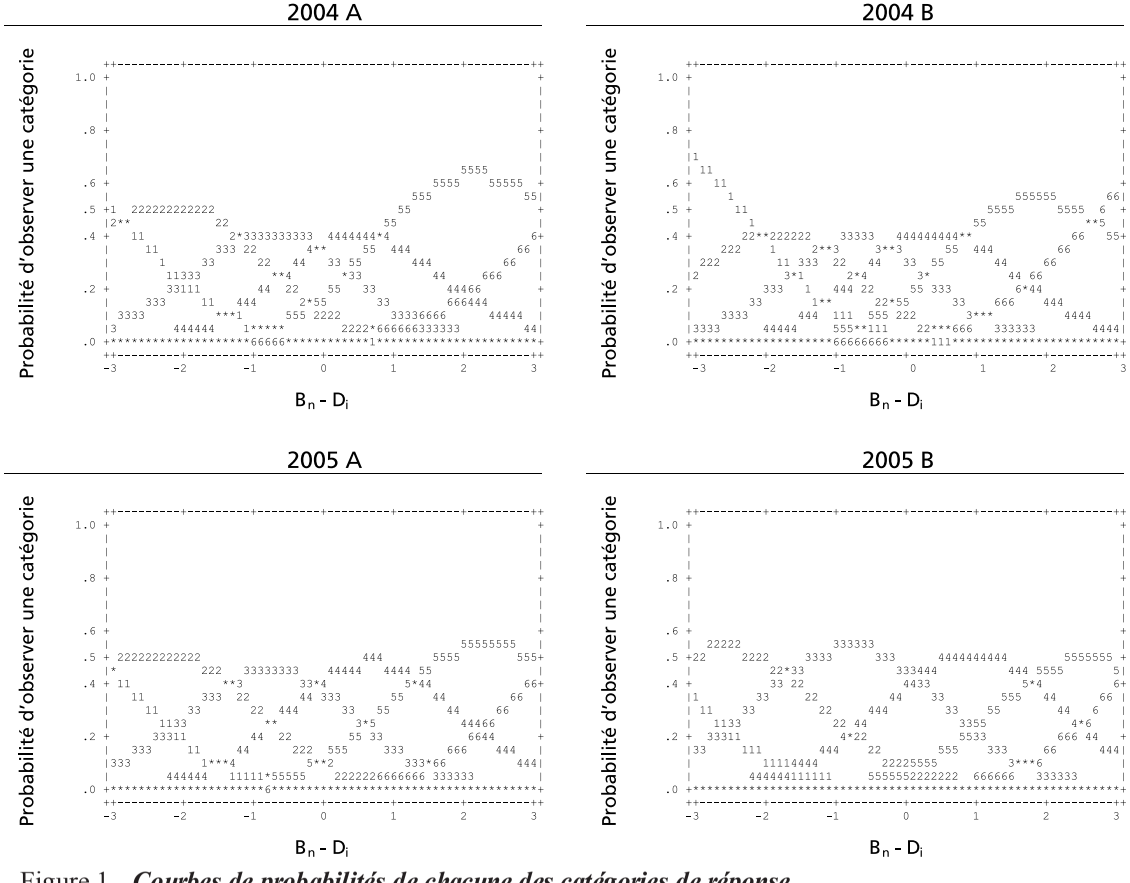

Figure 1. ***Courbes de probabilités de chacune des catégories de réponse pour les versions A et B des questionnaires 2004 et 2005***

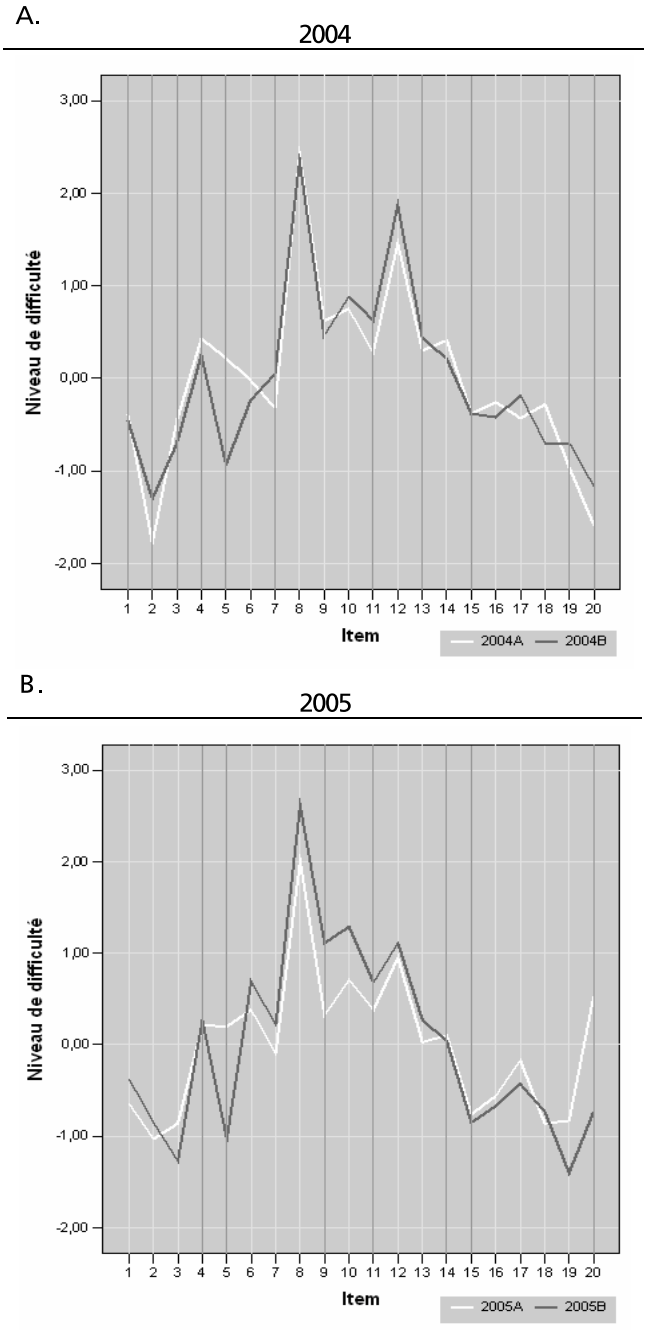

Figure 2. ***Position relative de la mesure estimée par le modèle pour le niveau d'endossement des items pour les versions A et B des questionnaires 2004 et 2005***

de l'échelle de réponses semblent exister (Linacre, 2006a). Il est possible que le nombre de catégories offertes aux répondants pour exprimer leur attitude était trop grand ou que la définition de la première et de la deuxième catégorie n'était pas suffisamment claire pour que les répondants puissent les distinguer facilement (de même pour la cinquième et la sixième catégorie). Aussi, il est possible que ces problèmes proviennent du mauvais ajustement trouvé pour certains items.

L'influence du choix des mots utilisés dans la formulation des items

Afin d'étudier l'influence du choix des mots utilisés dans la formulation des items, nous avons comparé la position relative des paramètres estimés par le modèle pour le niveau d'endossement des items des versions A et B des questionnaires, et ce, pour chacune des années. La figure 2A permet d'observer les différences entre les versions A et B du questionnaire 2004. Pour ces deux versions, les items les plus faciles à endosser sont les items 8 et 12, et les items les plus difficiles sont les items 20 et 2. Ces derniers portent sur les relations avec les parents. De plus, ces items font partie des 11 items pour lesquels la formulation a été modifiée. Il est également possible de voir que l'item 5 (item dont la formulation a été modifiée) est celui pour lequel la position relative diffère le plus entre les deux versions du questionnaire. La variation des items 2, 7, 11, 12, 18, 19 et 20 semble également notable. Parmi les items ayant été modifiés, les items 2, 7, 11, 13, 19 et 20 sont plus faciles à endosser dans la version B du questionnaire que dans la version A. Les items 1 et 15 occupent sensiblement la même position. Seuls les items 3, 4 et 5 sont plus difficiles à endosser dans la version B du questionnaire que dans la version A.

Pour l'année 2005, la figure 2B montre que les items 5 et 20 sont les items pour lesquels la différence entre les positions relatives de chacune des versions du questionnaire est la plus grande. La position de l'item 9 varie également un peu plus que les autres. La formulation de ces trois items a été modifiée. Il faut d'ailleurs rappeler que la formulation de l'item 20 a été complètement changée en 2005 et que les deux versions de cet item ne sont pas considérées *a priori* comme équivalentes. Des différences notables entre les positions relatives de presque tous les items peuvent être observées, sauf pour les items 2, 4, 12, 14, 15, 16 et 18 qui semblent assez stables d'une version à l'autre du questionnaire. Parmi ces derniers, trois items, les items 4, 14 et 15 sont des items pour lesquels la formulation a été modifiée. Les positions des autres items dont la formulation a été modifiée varient peu entre les deux versions du questionnaire.

Afin d'avoir une idée plus précise sur l'ampleur de ces variations et dans le but de comparer nos résultats à ceux de Schuman et Kalton (1985), les écarts entre les positions relatives des items des versions A et B des questionnaires ont été convertis en pourcentage. La figure 3 présente un diagramme en bâton dans lequel la hauteur des bâtons représente la valeur de la différence entre la mesure estimée par le modèle pour le niveau de difficulté des items de la version B par rapport à ceux de la version A. Les bâtons au-dessus de l'axe horizontal correspondent aux items qui sont plus faciles à endosser dans la version B que dans la version A. Les lignes de référence qui ont été ajoutées déterminent un intervalle de plus ou moins 5 % de variation entre les deux versions du questionnaire par rapport à l'échelle en six points utilisée.

Ainsi, la figure 3A révèle que la valeur de la mesure estimée par le modèle pour le niveau de difficulté de l'item 5 est 19,2 % plus élevée dans la version 2004B que dans la version 2004A. Il est donc beaucoup plus difficile d'endosser l'item 5 dans la version 2004B que dans la version 2004A. De plus, deux items pour lesquels la formulation n'a pas été changée ont une variation supérieure à 5 % (items 12 et 18). Parmi les items pour lesquels la formulation a été modifiée, cinq ont une variation supérieure à 5 % (items 2, 5, 7, 11 et 20), et cinq ont une variation inférieure à ce seuil (items 1, 3, 4, 13 et 19). Enfin, la variation de l'item 15 est nulle. À l'exception de l'item 5, la variation est toujours inférieure à 10 %. Pour l'année 2005, la figure 3B permet d'observer que la valeur de la mesure estimée par le modèle pour le niveau de difficulté de l'item 5 est 20,5 % plus élevée dans la version B que dans la version A. De même, la valeur estimée pour l'item 20 (item complètement différent dans les deux versions) est 21 % plus élevée dans la version B que dans la version A. Six items pour lesquels la formulation des items n'a pas été modifiée ont une variation supérieure à 5 % (items 3, 6, 7, 9, 11 et 19). Cette variation n'excède toutefois pas les 10 %. Parmi les items pour lesquels la formulation a été modifiée, trois items ont une variation supérieure à 5 % (items 5, 9 et 20). Dans les trois cas, la variation est même supérieure à 10 %.

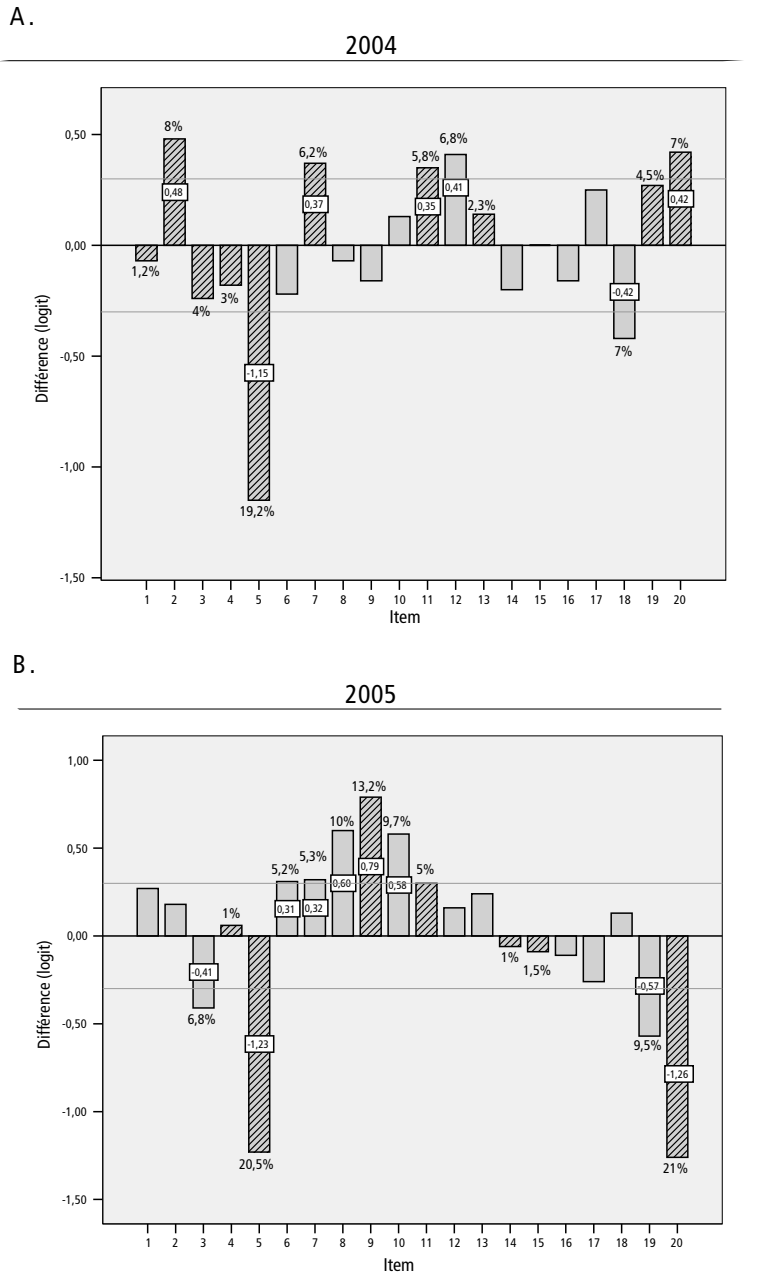

Figure 3. ***Diagrammes en bâton représentant la valeur des différences entre les positions relatives des mesures estimées par le modèle pour le niveau d'endossement des items entre les versions A et B des questionnaires 2004 et 2005***

Enfin, l'influence du choix des mots peut également avoir un impact sur les points d'ancrage de l'échelle de réponse. Le diagramme de gauche de la figure 4 révèle que, de façon générale, les points d'ancrage des items de la version B du questionnaire 2004 sont un peu plus comprimés que ceux des items de la version A. De plus, les différences les plus grandes entre les points d'ancrage des versions A et B se trouvent généralement aux extrémités des échelles, c'est-à-dire pour la deuxième et la sixième catégorie plus particulièrement. À l'inverse, les points d'ancrage des catégories intermédiaires (3, 4 et 5) sont plus rapprochés : l'intervalle entre ces catégories est généralement plus petit ou égal à 1 logit. L'analyse de ce diagramme ne permet pas d'établir un impact quelconque du choix des mots dans la formulation des items sur l'échelle de mesure associée aux catégories de réponses. En effet, les variations entre les points d'ancrage des deux versions du questionnaire sont assez similaires pour tous les items. De la même façon, le diagramme de droite de la figure 4 révèle que, de façon générale pour l'année 2005, les points d'ancrage des items de la version B du questionnaire sont un peu plus dilatés que ceux des items de la version A. La différence est toutefois assez petite. De plus, les points d'ancrage des troisième et quatrième catégories sont généralement un peu plus rapprochés que les autres. Ce rapprochement est plus prononcé dans la version A du questionnaire que dans la version B. Les différences les plus grandes entre les points d'ancrage des versions A et B peuvent être observées pour les items 5, 8, 9 et 20. Parmi ces items, seuls les items 5 et 20 sont des items pour lesquels la formulation a été modifiée. Encore une fois, il est difficile d'établir dans quelle mesure le choix des mots dans la formulation des items a pu avoir une influence sur l'échelle de mesure associée aux catégories de réponses.

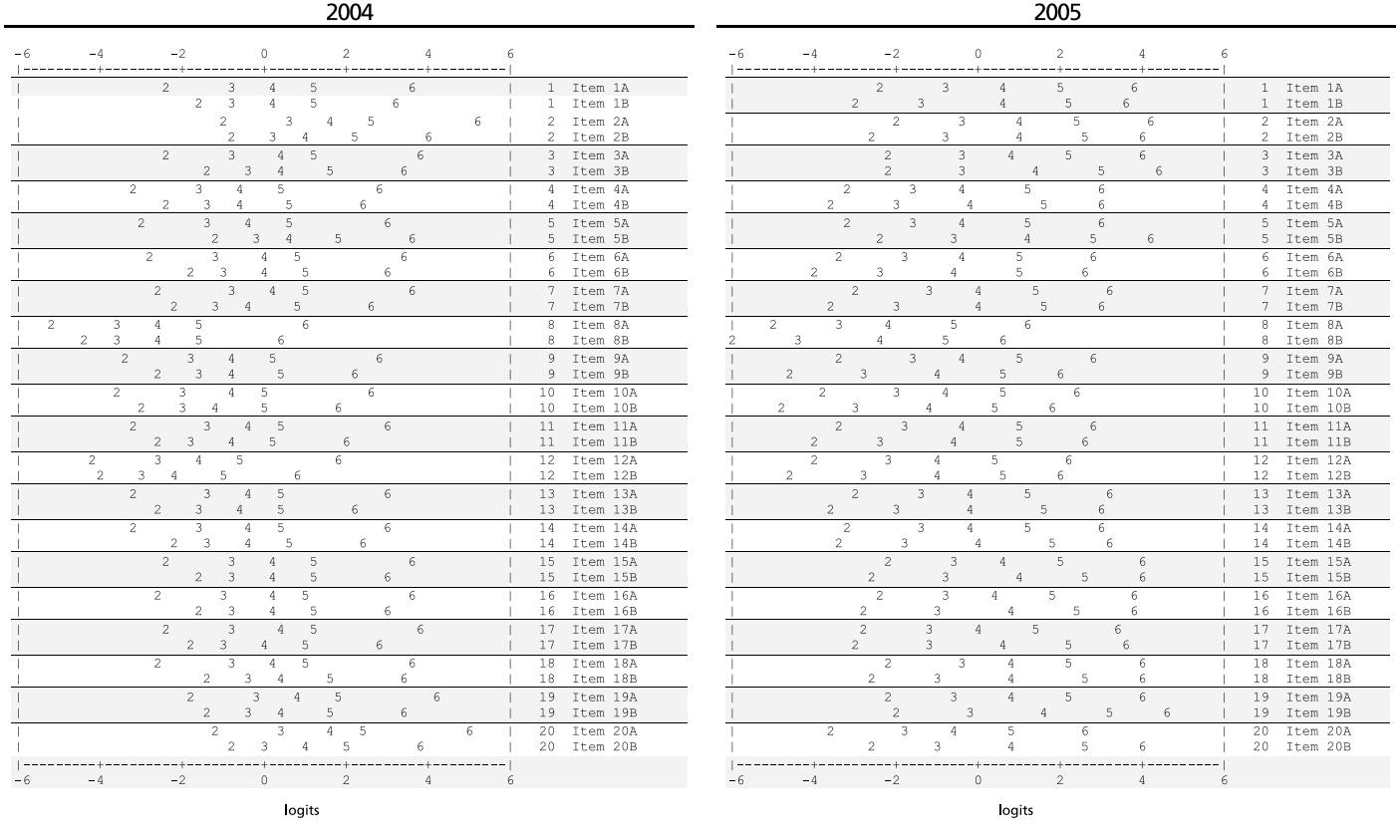

Figure 4. ***Diagramme de comparaison montrant la position des points d'ancrage estimés par le modèle pour chacun des items des versions A et B des questionnaires 2004 et 2005***

Discussion et conclusion

Tout d'abord, il est important de rappeler qu'il s'agit ici d'une recherche exploratoire avec une visée méthodologique. En effet, un des objectifs de l'étude consiste à déterminer dans quelle mesure le modèle de Rasch pour les catégories de réponses ordonnées et les outils pour le diagnostic de l'ajustement qui sont actuellement disponibles peuvent être mis à contribution dans l'étude des propriétés métriques des réponses récoltées avec des questionnaires d'enquête.

Soulignons également que la taille des groupes de personnes qui ont participé à cette recherche est restreinte. Malgré cela, les résultats obtenus dans cette étude laissent transparaître qu'il serait possible d'obtenir une certaine précision dans les résultats avec un échantillon d'au moins 50 personnes. Sur un total de 20 items, le nombre de ceux-ci dont le patron de réponse ne s'ajuste pas adéquatement au modèle est peu élevé. On en retrouve aucun pour la version B de 2004, seulement deux items pour la version 2004A et trois items pour les versions A et B de 2005. Selon Linacre (1994), un échantillon de cette taille permettrait effectivement d'obtenir des paramètres d'items assez stables (plus ou moins 1 logit) avec un intervalle de confiance à 99 %[2]. D'autres recherches portant précisément sur le lien entre le nombre de participants, les caractéristiques de la distribution des réponses pour chaque item et la qualité de l'ajustement pour les items, seraient toutefois nécessaires pour confirmer les résultats trouvés dans cette étude.

D'autre part, les résultats de cette étude révèlent que l'impact d'un changement de mots sur les réponses est variable et dépend des mots qui sont touchés par les modifications. En fait, ce ne serait pas tant la formulation (qui peut devenir plus ou moins précise selon le changement apporté) qui aurait un impact, mais le sens ou l'intention véhiculé par l'item. À titre d'illustration, les analyses des questionnaires 2004 ont montré qu'il n'existe aucune différence entre les positions des versions A et B de l'item 15. Pourtant, un changement de mot a été appliqué afin de rendre la formulation de la version B « Élaborer des outils pour l'évaluation formative » plus précise que celle de la version A « Construire des outils pour l'évaluation formative (exercices, devoirs, etc.) ». Or, il semble que la formulation plus ou moins précise de cet item ne change rien à l'intention que les répondants perçoivent par rapport à ce qui leur est demandé et n'influence donc pas les données recueillies.

D'autres résultats indiquent que la seule présence de certains mots dans la formulation d'un item peut jouer un rôle important sur l'intention que véhicule l'item et, par conséquent, sur les données recueillies, et ce, peu importe la formulation utilisée. L'exemple le plus saillant de ce constat concerne les items qui portent sur la relation avec les parents des élèves. Deux items de chacune des années étudiées portaient sur ce construit. Or, ces deux items sont toujours les plus difficiles à endosser par les personnes. Le même constat a été fait sur chacune des années du questionnaire depuis 2000. La particularité de toutes les formulations mises à l'essai est qu'au-delà des changements, elles contiennent toutes le mot *parents*. La présence du mot *parents* renvoie les étudiantes et les étudiants à leur programme d'étude et au constat qu'ils ne sont en aucune manière préparés à les rencontrer, à leur expliquer des choses au sujet de leur enfant, etc. On pourrait émettre l'hypothèse que la seule présence du mot *parents* et le contexte d'une rencontre avec ceux-ci suffiraient aux répondants pour qu'ils puissent se faire une opinion et qu'ils aient de la difficulté à endosser l'item qui leur est présenté.

Les analyses ont également fait ressortir que certains changements de mots pouvaient avoir un impact important sur les données recueillies. C'est le cas notamment pour l'item 5. Pour les deux années à l'étude, ce sont les deux mêmes formulations qui ont été utilisées dans les versions A et B. Ainsi, il est possible de suivre l'impact de la modification et de constater la stabilité des résultats. Pour les deux années, l'item propose des formulations où l'énoncé de la version A est moins général que celui de la version B : « Corriger les productions écrites des élèves » versus « Corriger la langue écrite des élèves ». Les différences observées pour cet item sont de 19,2 % en 2004 et de 20,5 % en 2005. La différence est assez marquée pour que l'on puisse avancer que deux choses différentes sont mesurées par les deux versions de l'item et qu'il s'agit bien de deux items différents. Le seul autre item qui a produit des différences aussi grandes est l'item 20 du questionnaire 2005 avec une différence de 21 % entre les versions A et B. Or cet item a été complètement changé et les deux versions de cet item n'étaient pas au départ considérées équivalentes. Les différences les plus grandes que cette étude a permis de mettre en lumière sont ces dernières et les variations sont donc de l'ordre de 20 %. L'item 9 du questionnaire 2005 a provoqué une différence de 13,2 %. Pour cet item, la formulation utilisée dans la version A, « Adapter mes activités d'enseignement aux caractéristiques des élèves », est plus précise que celle utilisée dans la version B, « Adapter mes activités d'enseignement ». Tous les autres items, qu'un changement de mot ait été appliqué ou non, ont provoqué des différences

de moins de 10 %. Ainsi, les résultats de la présente étude convergent dans le même sens que ceux obtenus par Schuman et Kalton en 1985. Il convient donc de s'interroger sur l'effet d'une telle différence pour la recherche. Est-ce que, comme le mentionnait Blais (1992), les variations peuvent être considérées suffisamment minces pour ne pas entraîner de différences majeures dans les résultats ? En effet, bien que la précision des résultats puisse en être affectée, jusqu'à quel point une variation de 10 % sur le continuum du niveau de l'endossement peut avoir des conséquences sur les résultats d'une enquête par questionnaire et, par suite, avoir un impact sur les résultats de recherche du chercheur ?

Par ailleurs, la différence de 21 % trouvée entre les deux versions de l'item 20 de 2005 permet d'apprécier l'utilité du modèle de mesure de Rasch. Dans la version A, l'item se lit « Guider les élèves dans leurs apprentissages » par rapport à « Discuter avec les élèves de leurs difficultés » pour la version B. Lorsqu'on examine les résultats, il est évident que ces deux items ne se situent pas au même endroit sur l'échelle du niveau d'endossement et qu'à ce titre, ce sont deux items qui contribuent différemment à l'échelle de mesure, plus différemment en fait que s'ils étaient situés exactement au même endroit. Ce sont donc deux items qui non seulement nous renverraient à des construits spécifiques différents, mais également deux items qui ne se dédoublent pas du point de vue de la mesure.

Enfin, les variations constatées pour les points d'ancrage des catégories de réponses estimés par le modèle étaient généralement du même ordre pour tous les items, c'est-à-dire autant pour les items dont la formulation avait été modifiée que pour les items qui n'avaient pas changé. Les résultats dans cette direction sont donc peu concluants et ne nous ont pas permis d'établir dans quelle mesure un changement de mots dans la formulation des items influence l'échelle de mesure associée aux catégories de réponses. En revanche, les résultats ont montré que certaines catégories de réponse sont parfois plus rapprochées que d'autres. Ainsi, les résultats de cette recherche ont permis d'observer que le nombre de points d'ancrage était peut-être trop élevé pour que les personnes puissent faire une distinction fine entre certains de ces points. Il serait intéressant, d'une part, de reprendre les analyses de cette recherche en utilisant le modèle *Partial Credit* pour voir si des conclusions différentes pourraient être dégagées et, d'autre part, de comparer les résultats pour deux versions A et B qui n'auraient pas, au départ, le même nombre de catégories.

En conclusion, on peut dire que lorsqu'on a comparé les résultats pour les deux versions du questionnaire avec comme objectif d'examiner l'impact du changement de mot sur les estimations de la position respectives des items, ce qu'on a observé nous porte à rester humble et à reconnaître que, dans l'ensemble, il y a peu de mouvements assez significatifs pour conclure à un impact important. En fait, il semble que dans la réalité, ce sont certains mots clés qui ont un impact et il n'est pas toujours aisé de déterminer quels sont ces mots clés sans avoir mis à l'essai des formulations différentes. À notre avis, le modèle de Rasch et les développements pour le diagnostic qui l'accompagnent se révèlent particulièrement adaptés pour le type d'instrument de collecte des données qui a été utilisé et le type de démarche emprunté. D'ailleurs, étant donnée la durée de la recherche (2000-2007), d'autres avenues, comme l'impact des étiquettes associées aux nombres dans les échelles de réponses (Blais & Grondin, 2008) ou le regroupement de catégories de réponses (Grondin & Blais, 2010), ont été explorées et ont fait l'objet de présentations dans différents forums scientifiques.

NOTES

1. Le questionnaire n'a pas été élaboré en vase clos. Ainsi, certaines modifications ont été apportées à la suite des discussions avec les responsables de l'opération d'évaluation des programmes au CFIM.
2. L'étude de Lee (1992) révèle qu'une différence d'un logit correspondrait à la différence, en termes de niveau d'habileté, entre les élèves d'une année scolaire à l'autre. Puisque la présente étude utilise deux groupes d'étudiants d'une même année universitaire, que ces deux groupes ont été formés par une distribution aléatoire des questionnaires et donc que ces deux groupes sont comparables *a priori*, et puisque les questionnaires utilisés portent sur l'opinion des étudiants au regard de leur programme de formation, les tailles d'échantillon utilisées dans cette étude nous apparaissent suffisantes pour effectuer des comparaisons.

RÉFÉRENCES

Abelson, R. P. (1992). Opportunities in survey measurement of attitudes. In J. M. Tanur (dir.), *Questions about questions* (pp. 173-176). NY: Russel Sage Foundation.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research and Perspectives, 9*(1), 95-104.

Blais, A. (1992). Le sondage. In B. Gauthier (dir.), *Recherche sociale : de la problématique à la collecte des données* (2e éd., pp. 367-398). Sillery, Québec : Presses de l'Université du Québec.

Blais, J.-G., & Ajar, D. (1992). Théorie des réponses aux items et modélisation. *Mesure et évaluation en éducation, 14*(4), 5-18.

Blais, J.-G., & Grondin, J. (2008, mars). *A study of the influence of labels associated with anchor points of Likert-type response scales in survey questionnaires.* Communication présentée lors du congrès de l'International Objective Measurement Workshop (IOMW), New York.

Blais, J.-G., & Grondin, J. (soumis). The influence of labels associated with anchor points of Likert-type response scales in survey questionnaires. *Journal of Applied Measurement.*

Blais, J.-G., Grondin, J., Loye, N., & Raîche, G. (2007, april). *A transverse study of items' wording impact with Rasch's rating scale model.* Communication présentée à la réunion annuelle de l'American Educational Research Association (AERA), Chicago, IL.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Borg, G. (1982). A category scale with ratio properties for intermodal and interindividual comparisons. In H.-G. Geissler & P. Petzold (dir.), *Psychophysical judgment and the process of perception* (pp. 25-34). Amsterdam: North-Holland Publishing Company.

Bradburn, N. M., & Sudman, S. (1979). *Improving interview method and questionnaire design: Response effects to threatenning questions in survey research*. San Francisco, CA: Jossey-Bass Publications.

Clark, H. H., & Schober, M. F. (1992). Asking questions and influencing answers. In J. M. Tanur (dir.), *Questions about questions* (pp. 15-48). NY: Russel Sage Foundation.

Cools, W., Hofmans, J., & Theuns, P. (2006). Context in category scales: Is "fully agree" equal to twice agree? *Revue européenne de psychologie appliquée, 56*, 223-229.

Curtis, D. D. (2003). *The influence of person misfit on measurement in attitude surveys*. Dissertation non publiée, Flinders University, Adelaide.

Curtis, D. D., & Boman, P. (2004, July). *The identification of misfitting response patterns to, and their influence on the calibration of, attitude survey instruments.* Communication présentée lors du 12e International Objective Measurement Workshop (IOMW), Cairns, Australia.

Dawes, R. M., & Smith, T. L. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (dir.), *Handbook of social psychology* (3e éd., vol. 1: Theory and method, pp. 509-566). New York: Random House.

Grondin, J., & Blais, J.-G. (2010, mai). *A Rasch analysis on collapsing categories in item's response scale of survey questionnaire: Maybe it's not one size fits all.* Communication présentée lors du congrès de l'American Educational Research Association (AERA), Denver, CO.

Groves, R. M., Fultz, N. H., & Martin, E. (1992). Direct questioning about comprehension in a survey setting. In J. M. Tanur (dir.), *Questions about questions* (pp. 49-61). NY: Russel Sage Foundation.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Plublications.

Jackson, T. R., & Popovich, N. G. (2006). The development, implementation, and evaluation of a self-assessment instrument for use in a pharmacy student competition. *American Journal of Pharmaceutical Education, 67*(2), article 57.

Kahneman, D., & Tversky, A. (1982). On the psychology of prediction. In D. Kahneman, P. Slovic & A. Tversky (dir.), *Judgment under uncertainty: Heuristics and biases* (pp. 48-68). Cambridge, UK: Cambridge University Press.

Krosnick, J. A., & Abelson, R. P. (1992). The case for measuring attitude strength in surveys. In J. M. Tanur (dir.), *Questions about questions* (pp. 177-203). NY: Russel Sage Foundation.

Lawton, G., Bhakta, B. B., Chamberlain, M. A., & Tennant, A. (2004). The Behçet's disease activity index. *Rheumatology, 43*(1), 73-78.

Lee, O. K. (1992). Variance in Mathematics and Reading across Grades. *Rasch Measurement Transcations, 6*(2), 222-223.

Li, M.-n. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21*(3), 215-231.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (1996). The Rasch model cannot be "disproved"! *Rasch Measurement Transactions, 10*(3), 512-514.

Linacre, J. M. (2006a). Aide en ligne du logiciel Winsteps Rasch Measurement (version 3.60.1). Consulté à partir de [www.winsteps.com].

Linacre, J. M. (2006b). Winsteps® Rasch Measurement (version: 3.60.1). Consulté à partir de [www.winsteps.com].

Linacre, J. M., & Wright, B. D. (1994). Dichotomous infit and outfit mean-square fit statistics. *Rasch Measurement Transactions, 8*(2), 350-360.

Ng, S. H., Pipe, M.-E., Beath, B., & Holton, D. (1999). Framing the problem: Effects of wording on children's statistical inferences. *Educational Psychology, 19*(4), 489-499.

Park, T. (2004). An investigation of an ESL placement test of writing using many-facet Rasch measurement. *Working papers in TESOL & applied linguistics, 4*(1).

Rohrmann, B. (2003). *Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data.* Melbourne, Australia: University of Melbourne.

Rugg, D. (1941). Experiments in wording questions: II. *Public opinion quarterly, 5*(1), 91-92.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology, 29*, 65-88.

Schuman, H., & Kalton, G. (1985). Survey methods. In G. Lindzey & E. Aronson (dir.), *Handbook of social psychology* (3e éd., vol. 1: Theory and method, pp. 635-697). New York: Random House.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93-105.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions, 10*(3), 516-517.

Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith Jr. & R. M. Smith (dir.), *Introduction to Rasch measurement: theory, models and applications* (pp. 73-92). Maple Grove, MN: JAM Press.

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*(1), 66-78.

Smith, R. M., & Suh, K. K. (2003). Rasch fit statistics as a test of the invariance of item parameter estimates. *Journal of Applied Measurement, 4*(2), 153-163.

Smith, T. W. (1980). The 75% solution: An analysis of the structure of attitudes on gun control, 1959-1977. *The Journal of Criminal Law & Criminology, 71*(3), 300-316.

Tejada, A. J. R., Gómez, A. G., García, J. L. P., & Meléndez, C. P. (2002). Two strategies for fitting real data to Rasch polytomous models. *Journal of Applied Measurement, 3*(2), 129-145.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*(3), 299-314.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Wang, W.-C., & Chen, C.-T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. *Educational and Psychological Measurement, 65*(3), 376-404.

Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith Jr. & R. M. Smith (dir.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 1-24). Maple Grove, MN: JAM Press.

Date de réception : 6 janvier 2009

Date de réception de la version finale : 16 novembre 2010

Date d'acceptation : 18 novembre 2010