Mesure et évaluation en éducation

Estimation des paramètres d'item et de sujet à partir du modèle de Rasch

Une étude comparative des logiciels BILOG-MG, ICL et R

Sébastien Béland, David Magis and Gilles Raîche

Volume 36, Number 1, 2013

URI: https://id.erudit.org/iderudit/1024466ar
DOI: https://doi.org/10.7202/1024466ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

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

Explore this journal

Cite this article

Béland, S., Magis, D. & Raîche, G. (2013). Estimation des paramètres d'item et de sujet à partir du modèle de Rasch : une étude comparative des logiciels BILOG-MG, ICL et R. *Mesure et évaluation en éducation*, *36*(1), 83–110. https://doi.org/10.7202/1024466ar

Article abstract

Item response theory (IRT) is a class of measurement models extensively used in education. Curently, many softwares are available (e.g. BILOG-MG) for the estimation of item and examinee parameters. Among these softwares, one must mention ICL and R, which are free and allow to produce to a large variety of analyses. The main objective of this study is to use the Rasch model to compare the quality of estimation of the difficulty and subject parameters. Here, we will compare item parameters through three software packages: BILOG-MG, ICL and the R package ltm. The demonstration will be twofold: we will make a simulation study and an analysis of an English proficiency test, as second language. Our results show that these softwares obtained similar parameters estimates, their main difference pertaining to their respective computation times.

Estimation des paramètres d'item et de sujet à partir du modèle de Rasch : une étude comparative des logiciels BILOG-MG, ICL et R

Sébastien Béland
Université du Québec à Montréal

David Magis
Université de Liège (Belgique)

Gilles Raîche
Université du Québec à Montréal

Mots clés : modèle de Rasch, paramètre de difficulté d'item, paramètre de sujet, BILOG-MG, R, ICL

La Théorie de la réponse aux items (TRI) est une classe de modèles de mesure très utilisée en éducation. À ce jour, de nombreux logiciels, tel BILOG-MG, sont disponibles afin de procéder à l'estimation des paramètres d'item et de sujet. Parmi ces logiciels, il ne faut pas négliger ICL et R qui sont gratuits et qui peuvent permettre de produire des analyses diversifiées. Cette étude a pour objectif de comparer la qualité d'estimation des paramètres selon une des modélisations issues de la TRI : le modèle de Rasch. Pour ce faire, nous comparons les estimateurs du paramètre de difficulté et de sujet selon trois logiciels : BILOG-MG, ICL et la librairie ltm, disponible sous le logiciel R. Nous procédons à une analyse par simulation informatique et, dans un second temps, nous analysons un test de classement en anglais, langue seconde. Les résultats démontrent que les logiciels étudiés permettent d'obtenir des estimateurs des paramètres similaires, la différence principale entre ces logiciels étant leur temps d'exécution des procédures d'estimation.

Key words: Rasch model, item difficulty parameter, subject parameter, BILOG-MG, R, ICL

Item response theory (IRT) is a class of measurement models extensively used in education. Curently, many softwares are available (e.g. BILOG-MG) for the estimation of item and examinee parameters. Among these softwares, one must

mention ICL and R, which are free and allow to produce to a large variety of analyses. The main objective of this study is to use the Rasch model to compare the quality of estimation of the difficulty and subject parameters. Here, we will compare item parameters through three software packages: BILOG-MG, ICL and the R package ltm. The demonstration will be twofold: we will make a simulation study and an analysis of an English proficiency test, as second language. Our results show that these softwares obtained similar parameters estimates, their main difference pertaining to their respective computation times.

Palavras-chave: modelo Rasch, parâmetro de dificuldade de item, parâmetro de sujeito, BILOG-MG, R, ICL

A teoria de resposta aos itens (TRI) é uma classe de modelos de medida muito utilizada em educação. Atualmente, muitos softwares, como BILOG-MG, estão disponíveis para a estimação dos parâmetros de item e de sujeito. Entre estes softwares, não se deve negligenciar o ICL e R, os quais são gratuitos e podem permitir análises diversificadas. Este estudo tem por objetivo comparar a qualidade de estimação dos parâmetros segundo uma das modelizações da TRI: o modelo Rasch. Para isso, comparamos os estimadores do parâmetro de dificuldade e de sujeito segundo três softwares: BILOG-MG, ICL e a biblioteca ltm disponível no software R. Procedemos a uma análise por simulação informática e, num segundo tempo, analisamos um teste proficiência em inglês como segunda língua. Os resultados demonstram que os softwares estudados permitem obter de estimadores de parâmetros similares, sendo que a diferença principal entre estes softwares é o tempo de execução dos procedimentos de estimação.

Note des auteurs – Toute correspondance peut être adressée comme suit : Sébastien Béland, Département d'éducation et de pédagogie, Faculté des sciences de l'éducation, Université du Québec à Montréal, 1205, rue Saint-Denis, bureau N-6565, Montréal (QC), Canada H2X 3R9, télécopieur : 514-987-4608, ou par courriel à l'adresse suivante : [beland.sebastien@gmail.com].

Introduction

Bien que les premières épreuves de classement d'individus remontent à la Chine de 1115 av. J.-C. (Bertrand & Blais, 2004), l'histoire de la mesure débute à la fin du 19e siècle avec les études d'auteurs aussi connus que Francis Galton, Alfred Binet et Charles Spearman (Hambleton & Swaminathan, 1985 ; Sijtsma & Junker, 2006). Ces travaux pionniers s'étendent surtout jusqu'aux années 1960 et sont, aujourd'hui, connus sous le vocable de Théorie classique des tests.

Dans le cadre de cet article, les auteurs s'intéressent plus particulièrement à une approche qui s'est développée à la suite de la Deuxième Guerre mondiale : la Théorie de la Réponse aux Items (TRI ; Birnbaum, 1968 ; De Ayala, 2009 ; Hambleton & Swaminathan, 1985 ; Lord, 1980). C'est surtout la recherche de modèles respectant le principe d'invariance de la mesure et le développement de la micro-informatique qui ont permis à cette théorie de devenir populaire. Les modélisations issues de la TRI sont d'ailleurs fort utiles pour permettre d'analyser les données issues de grandes enquêtes internationales telle que le Programme for International Student Assessment (PISA).

À ce jour, plusieurs logiciels sont disponibles pour procéder à l'analyse de données selon les modélisations issues de la TRI. Par exemple, il est possible de se référer à BILOG-MG, RUM2030, Xcalibre, PARSCALE, PARAM-3PL, WINSTEPS/BIGSTEPS ou ConQuest. Malheureusement, ces logiciels souffrent tous d'un double problème : le coût, d'une part, et la documentation de soutien à l'utilisateur qui est relativement limitée.

L'avènement du web 2.0[1] a permis une certaine démocratisation de la programmation statistique. À ce titre, il est possible de penser au logiciel *R Development Core Team* (2012), qui est gratuit, fort documenté et permet de produire une panoplie d'analyses basées sur la TRI. Développé dans les laboratoires de Bell, ce logiciel de type *open source* renferme aujourd'hui un grand nombre de modèles statistiques et des fonctions graphiques flexibles de qualité supérieure. Ce logiciel est toutefois moins utilisé par les chercheurs en éducation et en sciences humaines. Selon les auteurs, au moins deux raisons peuvent expliquer cela : son interface est moins conviviale que celle d'autres logiciels comme SPSS et son utilisation requiert des connaissances de base

en programmation. Cependant, au regard des analyses associées à la TRI, cela ne devrait pas être une limite sérieuse. Les autres logiciels dédiés à ces modélisations souffrent presque tous de ces deux mêmes problèmes.

À la connaissance des auteurs, aucune recherche publiée ne s'est encore attardée à étudier la qualité de librairies R permettant d'analyser des données à l'aide de la TRI. Dans le cadre de cet article, les auteurs ont choisi une perspective comparative en évaluant les logiciels BILOG-MG, ICL et la librairie ltm (disponible dans le logiciel R) pour estimer les paramètres d'item (par ex., le maximum de vraisemblance marginale) et de sujet (par ex., le maximum de vraisemblance) sous le modèle de Rasch. Pour ce faire, dans un premier temps, une simulation informatique en supposant des conditions décrites à la section portant sur la méthodologie a été utilisée. Dans un deuxième temps, les données du test de classement en anglais, langue seconde, au collégial (TCALS-II ; Laurier, Froio, Pearo, & Fournier, 1998) sont analysées.

La section suivante présente la modélisation à réponse dichotomique de Rasch ainsi que les logiciels utilisés pour estimer les paramètres de ce modèle. Ensuite, le cadre méthodologique sera décrit, puis les résultats des analyses. Finalement, une discussion et une brève conclusion seront présentées.

Cadre théorique

Le concept intégrateur de la TRI consiste à utiliser des données manifestes, représentées par des réponses à des items portant sur un contenu disciplinaire précis, afin d'obtenir une information à l'égard d'un trait latent (noté θ, dans la figure 1) non immédiatement mesurable. Comme Benzécri (1973) le rapportait : Piaget énonçait que « la grande difficulté de la psychologie est l'absence d'unité de mesure » (p. 23). En éducation aussi, l'habileté d'un étudiant n'est pas directement observable (Hulin, Drasgow, & Parson, 1983). L'évaluation du niveau d'habileté ne se prête pas directement à la mesure : il faut passer par l'intermédiaire d'un outil, d'un test, pour recueillir des données ou des observations manifestes (soit les items contenus au sein d'un test) afin d'approcher une mesure de ce niveau d'habileté. Or, le test n'est qu'une des multiples occasions que possède un individu de manifester son habileté (Lord, 1952). En accord avec ce dernier énoncé, Hambleton et Swaminathan (1985) vont un peu plus loin : cette habileté peut être cognitive, porter sur des caractéristiques de la personnalité, être une compétence de base, etc. De plus, le niveau de cette habileté ne serait pas inné : il peut changer dans le temps (Rentz

& Bashaw, 1977). En résumé, il peut être avancer que l'habileté peut être bien des choses qu'on ne peut observer (par ex., l'habileté en mathématiques ou l'habileté en lecture), mais sur laquelle on peut inférer certains renseignements pertinents à partir de données qui, elles, sont manifestes.

La figure 1 illustre schématiquement l'idée précédente. Le trait latent θ, représenté par un cercle, est considéré comme la cause des réponses aux quatre items, notés x_1 à x_4 et représentés par des carrés ; ceci explique pourquoi ce sont les flèches qui relient le trait latent aux quatre items.

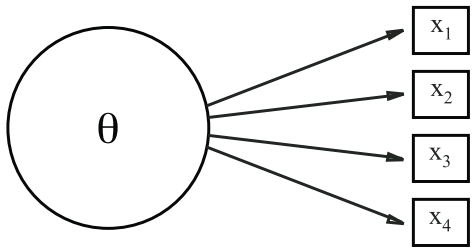

Figure 1. ***Données manifestes et trait latent***

En éducation, il est souhaité, par exemple, mesurer le niveau d'habileté (considérée comme le trait latent θ) d'un étudiant à partir des bonnes ou des mauvaises réponses manifestes qu'il a fournies à une série d'items. Ainsi, un item à choix multiples contient généralement une bonne et plusieurs mauvaises options où un score de 1 équivaut à une bonne réponse et un score de 0 à une mauvaise réponse ; voilà pourquoi il s'agit d'une modélisation à réponse dichotomique. Pour illustrer cela, imaginons l'item suivant qui permet la manifestation de l'habileté en mathématiques :

Félix a 3 pommes et 2 carottes. Anne-Sophie, de son côté, a 1 poire et 2 oignons. Combien de fruits ont Félix et Anne-Sophie?

i) 2 fruits
ii) 3 fruits
iii) 4 fruits

Figure 2. ***Exemple d'item***

Dans ce cas-ci, il existe une seule bonne option (iii) et deux mauvaises. Ainsi, un étudiant qui aurait répondu « iii » obtiendrait un score de 1 alors qu'un étudiant répondant « i » ou « ii » obtiendrait le score 0. Voyons maintenant comment analyser les matrices de données à l'aide du modèle de Rasch (1960).

Le modèle à réponses dichotomiques de Rasch

Il existe plusieurs modélisations permettant d'analyser des matrices de données à réponses dichotomiques. Par exemple, le modèle logistique à trois paramètres (Birnbaum, 1968) est intéressant puisqu'il permet de généraliser les modèles de base de la TRI. Cependant, dans le cadre de cet article, les auteurs s'intéresseront seulement au modèle de Rasch (1960). Une première raison est qu'avec l'étude de plus petits échantillons de sujets, comme dans l'exemple utilisé dans cette recherche, une modélisation comportant moins de paramètres à estimer permet d'obtenir des estimateurs d'une précision satisfaisante. Ce n'est généralement pas le cas avec les modélisations comportant plus de paramètres, surtout si ces derniers risquent d'être fortement corrélés. De plus, dans le cas de la modélisation à trois paramètres, l'estimation du paramètre de pseudo-chance nécessite un grand nombre de sujets qui affichent un niveau d'habileté faible : une situation qui est difficile à respecter si l'échantillon de sujets est trop petit. Dans ce contexte, plusieurs auteurs ont démontré que le modèle de Rasch respecte de façon raisonnable le principe d'invariance des items et des sujets pour différents tests et différents groupes de répondants (Forsyth, Sarsangjan, & Gilmer, 1981) et qu'il est possible d'obtenir des estimations stables des paramètres avec des échantillons de petite taille (Wright & Stone, 1979).

Mathématiquement, dans le modèle de Rasch, on calcule la probabilité $P_i(\theta)$ qu'un sujet de niveau d'habileté θ obtienne une bonne réponse à l'item i à partir de l'équation suivante :

$$P_i(\theta) = P(x_i = 1 \,|\, \theta, b_i) = \frac{\exp[(\theta - b_i)]}{1 + \exp[(\theta - b_i)]} \qquad (1)$$

où il est possible de retrouver le paramètre de difficulté d'item b_i. Ce modèle reconnaît que la probabilité de répondre correctement à l'item augmente avec le niveau d'habileté de l'étudiant et diminue lorsque le niveau de difficulté de l'item augmente. De plus, deux grandes conditions d'application doivent être vérifiées pour que le modèle de Rasch puisse être utilisé. Premièrement, il doit y avoir indépendance locale entre les items d'un test. Ainsi, pour un niveau

d'habileté θ fixé, les réponses fournies à deux items (x_1 et x_2) doivent être indépendantes lorsque le niveau d'habileté est fixé (et donc non corrélées), comme l'indique le théorème multiplicatif en probabilité suivant :

$$P(x_1 \wedge x_2 \mid \theta) = P(x_1 \mid \theta)P(x_2 \mid \theta) \quad (2)$$

Deuxièmement, il ne peut y avoir qu'une seule habileté mesurée dans l'épreuve d'évaluation : c'est l'hypothèse d'unidimensionnalité du trait latent. Par exemple, dans le cadre d'un test de classement en anglais, langue seconde, seule l'habileté, en anglais, doit être testée.

L'estimation des paramètres (item et sujet)

Les valeurs numériques des paramètres présentés à l'équation (1) doivent être disponibles afin que l'on puisse calculer les probabilités $P_i(\theta)$. La valeur du paramètre d'item b_i peut être estimée à partir de plusieurs méthodes qui ont déjà été exposées en détail par Hambleton et Swaminathan (1985) ainsi que par Baker et Kim (2004). Citons à titre d'exemple la méthode du maximum de vraisemblance conjointe, du maximum de vraisemblance marginale, du maximum de vraisemblance conditionnelle ou les méthodes bayésiennes utilisant les chaînes de Markov Monte Carlo (*Monte Carlo Markov Chain*). Dans cet article, l'attention est porté aux logiciels exploitant l'approche par maximum de vraisemblance marginale, approche généralement la plus utilisée. L'estimateur du niveau d'habileté θ, pour sa part, peut être obtenu à partir des méthodes d'estimation par vraisemblance maximale bayésienne ou non, par maximum de vraisemblance pondérée (*weighted maximum likelihood*) ou encore par espérance mathématique *a posteriori*. C'est la méthode par maximum de vraisemblance qui sera retenue dans le cadre de cet article.

Les logiciels étudiés

Même s'il existe plusieurs logiciels permettant d'analyser des données à l'aide du modèle de Rasch, nous nous concentrerons sur des logiciels relativement différents : BILOG-MG, ICL et la librairie ltm disponible dans le logiciel R. La justification de ce choix de logiciels est quadruple. Premièrement, ces logiciels émergent de philosophies différentes : BILOG-MG est un logiciel commercial alors que R et ICL sont gratuits et disponibles en logiciels libres (*open source*). Deuxièmement, le logiciel BILOG-MG est un environnement fermé alors que R est ouvert à tous les développeurs intéressés. Troisièmement, les librairies du logiciel R nécessitent une validation des résultats obtenus.

Les auteurs compareront donc les résultats obtenus dans ltm à ceux d'un logiciel ayant été éprouvé par les chercheurs et l'industrie : BILOG-MG. Pour pousser l'analyse un peu plus loin, les auteurs incluent dans la comparaison le logiciel ICL retrouvé aussi dans la littérature. Quatrièmement, les auteurs se sont concentrés sur des logiciels qui utilisent tous la même méthode d'estimation des paramètres d'item : le maximum de vraisemblance marginale. Les résultats obtenus ne seront donc pas affectés par une différence méthodologique d'estimation des paramètres d'item et d'habileté selon les logiciels. Malheureusement, seule la librairie ltm utilise actuellement cette méthode d'estimation, ce qui a donc limité le nombre de librairies de R à inclure dans cette étude.

Le logiciel BILOG-MG

BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) est fort probablement l'un des logiciels les plus connus et les plus utilisés dans le domaine de la TRI. Ce programme offre à l'utilisateur une panoplie de modèles (par ex., le modèle de Rasch, les modèles logistiques à un, deux et trois paramètres) permettant d'analyser des données à réponses dichotomiques en un temps de calcul généralement court. L'utilisateur peut réaliser l'estimation des paramètres d'item à l'aide du maximum de vraisemblance marginale et le logiciel intègre des mesures de l'adéquation de l'estimation des items au modèle de Rasch. Le paramètre θ, de son côté, est estimé à l'aide de méthodes telles que le maximum de vraisemblance, le maximum de vraisemblance *a posteriori* et la méthode d'espérance mathématique *a posteriori*. Le lecteur intéressé à en savoir plus sur BILOG-MG est invité à consulter l'article de Rupp (2003).

Le logiciel ICL

Développé par Hanson (2002), le logiciel *IRT Command Language* (ICL) version 0.020301 permet d'analyser des données à réponses dichotomiques et polytomiques selon plusieurs modèles issus de la TRI. Ce logiciel libre et gratuit permet d'utiliser deux types d'estimateurs des paramètres d'item : le maximum de vraisemblance marginale et le maximum de vraisemblance marginale bayésienne. De leur côté, les paramètres de sujet peuvent être estimés à l'aide de méthodes telles que le maximum de vraisemblance et sa version bayésienne. De plus, il est possible de vérifier l'adéquation des données, indépendamment du modèle de réponse à l'item utilisé.

Bien que moins connu, ce logiciel a tout de même déjà fait l'objet de quelques études. Par exemple, Jurich et Goodman (2009) ont déjà démontré que ICL et PARSCALE, un logiciel utilisé dans l'industrie, présentent des estimations équivalentes pour les paramètres d'item et d'habileté. De plus, Mead, Morris, et Blitz (2007) ont soulevé le fait que ICL présentait plus de flexibilité et d'options que BILOG-MG lorsque venait le temps de procéder à l'estimation de données à l'aide du modèle à trois paramètres.

La librairie ltm sous le logiciel R

Le logiciel R permet de procéder à l'analyse de données à réponses dichotomiques à l'aide du modèle de Rasch (1960). Par contre, il n'y a pas d'évaluation systématique des librairies. Seuls quelques articles ont discuté des applications de la TRI avec R. Par exemple, Rizopoulos (2006, 2012) a présenté la librairie ltm alors que Mair et Hatzinger (2007a, 2007b) ont présenté la librairie eRm (Mair, Hatzinger, & Maier, 2010) ou Weeks (2010, 2011) la librairie plink. De leur côté, De Boeck et al. (2011) ont montré comment l'utilisation de la librairie lmer permet de procéder à des analyses étendues de certains modèles traditionnels de la TRI. Enfin, d'autres articles traitent d'applications précises de la TRI. Par exemple, Magis, Béland, Tuerlinckx, et De Boeck (2010) ont présenté la librairie difR (Magis, Béland, & Raîche, 2012) qui permet d'utiliser plusieurs méthodes pour procéder à des analyses du fonctionnement différentiel d'items. Plus récemment, Magis et Raîche (2011, 2012) ont développé une librairie permettant de réaliser la simulation de tests adaptatifs, soit catR.

Dans le cadre de cette étude, les auteurs ont sélectionné une seule librairie pour procéder aux comparaisons de logiciels : ltm (Rizopoulos, 2012), version 0.9-9. Cette sélection a été basée sur le fait que la description de la librairie a été publiée dans une revue arbitrée scientifiquement (*Journal of Statistical Software*) et qu'elle est la seule à estimer les paramètres d'item par maximum de vraisemblance marginale, retenue dans cette étude. De plus, elle permet d'estimer le niveau d'habileté des répondants à l'aide de méthodes telles que le maximum de vraisemblance *a posteriori* et la méthode d'espérance mathématique *a posteriori*.

Dans ce qui suit, nous étudierons l'estimation des paramètres d'item et de sujet issus du modèle de Rasch en comparant trois logiciels : BILOG-MG, ICL et la librairie ltm (disponible dans le logiciel R). Pour ce faire, les auteurs

procéderont en deux temps : premièrement ils conduiront une étude de simulation par ordinateur et, ensuite avec des données réelles, une étude d'un test de classement en anglais, langue seconde.

Méthode

Les logiciels ont été sélectionnés parce qu'ils sont relativement différents et qu'ils utilisent certaines méthodes d'estimation communes. Par contre, les paramètres de difficulté ne sont pas tous représentés sous la même métrique au sein de ces logiciels. Pour produire des analyses comparables, les auteurs ont été dans l'obligation de procéder à deux manipulations supplémentaires.

Premièrement, la librairie ltm propose des estimations où $\log(\theta + b_i)$ est utilisé comme dans un contexte de régression linéaire alors que les logiciels ICL et BILOG-MG offrent des estimations basées sur la formule classique $\log(\theta - b_i)$, comme à l'équation (1). Ainsi pour ICL et BILOG-MG, b_i est égal à l'inverse sous l'addition de la valeur du même paramètre avec ltm. Le signe des estimations de ltm a donc été inversé pour les ramener à la convention utilisée par les autres logiciels. Deuxièmement, le logiciel BILOG-MG force les paramètres de difficulté à être de moyenne nulle, alors que cette contrainte n'apparaît pas avec les logiciels ICL et ltm. Pour obtenir des estimations comparables, les paramètres de difficulté obtenus avec ICL et ltm ont donc aussi été recentrés pour qu'ils soient de moyenne nulle. Les auteurs s'inspirent, dans le cas du modèle de Rasch, de la méthode de mise à l'échelle de type « *mean/mean* » qui est discutée dans Kolen et Brennan (2004).

Les comparaisons sont effectuées sur la base de deux études : une simulation par ordinateur et une analyse des données réelles obtenues lors de l'administration d'un test de classement en anglais, langue seconde, au collégial, soit le TCALS-II. De plus, les analyses ont été produites à l'aide de la librairie irtoys (Partchev, 2011) du logiciel R, qui permet d'effectuer tous les calculs et permet aussi les appels externes aux logiciels BILOG-MG, ICL et à la librairie ltm au sein de l'environnement d'analyse du logiciel R. À cet effet, il est important de comprendre que irtoys n'est qu'une plateforme à partir de laquelle il est possible d'importer tous les résultats produits par BILOG-MG, ICL et ltm.

Étude 1 : Simulation par ordinateur

Une simulation de type Monte Carlo a été effectuée pour comparer les paramètres de difficulté et de sujet à l'aide de ltm, ICL et BILOG-MG et vérifier s'ils présentent des résultats similaires dans différentes situations.

Les simulations ont été menées selon un plan factoriel complètement croisé avec quatre facteurs :

1) la longueur du test ;
2) la taille de l'échantillon d'étudiants ;
3) la distribution de probabilité du niveau d'habileté des étudiants ; et
4) la méthode d'estimation des paramètres (item et sujet) du modèle de Rasch.

Il est possible de supposer que la précision des estimateurs des paramètres d'item augmente avec la taille de l'échantillon d'étudiants. Le choix d'une distribution particulière pour les niveaux d'habileté pourrait avoir un impact direct sur les estimations des paramètres, notamment avec les méthodes de maximum de vraisemblance marginale.

Quatre longueurs de test ont été considérées : 20 items, 40 items, 60 items et 80 items. De même, quatre tailles d'échantillons d'étudiants ont été sélectionnées, correspondant à des groupes de 100, 500, 1 000 et 5 000 étudiants. En ce qui concerne les distributions de probabilité du niveau d'habileté, trois distributions ont été choisies : la loi normale standard N(0,1), la loi uniforme U(-2,2) sur l'intervalle [-2,2], et la loi bêta(5,17) sur l'intervalle [0,1]. La première loi est un choix classique en TRI : elle suppose une distribution symétrique autour de la valeur centrale zéro et ramène l'écart type à 1. La loi uniforme permet de modéliser une répartition uniforme des niveaux d'habileté des étudiants sur le continuum du trait latent, rapporté ici à l'intervalle [-2,2]. Enfin, la loi bêta(5,17) permet de modéliser une distribution asymétrique des niveaux d'habileté, avec une dissymétrie plus marquée à gauche et donc, une proportion plus importante de faibles niveaux d'habileté. Les valeurs θ générées selon cette loi sont ensuite centrées et réduites, de manière à obtenir des niveaux d'habileté centrés sur zéro et dispersés de façon dissymétrique avec un écart type de 1.

Il est à noter finalement que les paramètres de difficulté ont été générés selon une loi normale standard N(0,1). Il n'y a pas de raison *a priori* d'envisager un impact du choix de cette distribution sur la qualité de l'estimation des méthodes. C'est pourquoi d'autres distributions de probabilité pour les paramètres d'item n'ont pas été envisagées.

La génération des réponses des sujets aux items a été réalisée de la façon suivante. Pour un sujet de niveau d'habileté θ, la réponse x_i de ce sujet à l'item i (de niveau de difficulté b_i) a été générée aléatoirement selon une loi de Bernoulli dont la probabilité de succès est égale à $P_i(\theta)$. De manière équivalente, un nombre a été choisi aléatoirement sur l'intervalle [0,1] et comparé à la valeur $P_i(\theta)$. La réponse x_i a alors pris la valeur 1 si ce nombre était inférieur à $P_i(\theta)$, sinon, x_i a pris la valeur 0.

Cette étude a donc considéré 48 situations différentes obtenues en croisant les quatre longueurs de tests avec les quatre tailles d'échantillons et les trois distributions de probabilité du niveau d'habileté. Pour chacune des 48 situations, 100 jeux de paramètres d'item ont été générés et 100 ensembles de données (un par jeu de paramètres d'item) ont été créés. Les trois logiciels, ltm, ICL et BILOG-MG, ont permis de calculer les paramètres (item et sujet) avec chaque ensemble de données.

La comparaison finale des résultats obtenus par chacun des logiciels est basée sur trois statistiques :

1) la racine du carré moyen résiduel (RCMR) des paramètres d'item (*rootmean square error*) ;
2) la corrélation entre les paramètres estimés et les paramètres réels ; et
3) le temps de calcul requis pour estimer les paramètres d'item.

Le carré moyen résiduel (CMR) est la moyenne des carrés des écarts entre paramètres réels et estimés ; on prend en général la racine carrée du CMR pour rester sur la même échelle de mesure que le biais. De petites valeurs du RCMR indiquent que la méthode d'estimation est à la fois peu ou pas biaisée et donc précise. Enfin, la corrélation permet de mesurer globalement l'adéquation linéaire entre les paramètres d'item (estimés et réels) tandis que le temps de calcul reflète l'efficacité pratique d'une méthode en termes d'effort de calcul. Il est à noter que ces valeurs sont résumées par la suite à partir de leur moyenne pour les 100 répétitions de jeux de données.

Dans un deuxième temps, les auteurs ont souhaité voir comment l'estimation du paramètre d'item pouvait aussi affecter l'estimation du paramètre de sujet. Pour ce faire, les auteurs ont procédé en utilisant séparément les paramètres d'item estimés par les logiciels ltm, BILOG-MG et ICL, ce qui donne trois jeux de paramètres d'item par ensemble de données. Ensuite, ils ont procédé à l'estimation des paramètres de sujet selon deux techniques : le maximum de vraisemblance (Lord, 1980) et le maximum de vraisemblance pondérée (Warm, 1989). Il est important de comprendre que ces estimations ont été réalisées à l'aide de la librairie catR (Magis & Raîche, 2012), du logiciel R, plutôt qu'à l'aide des logiciels BILOG-MG, ICL et ltm. Ainsi, ces derniers n'ont servi qu'à la calibration des items. Finalement, les résultats de l'estimation des paramètres de sujet sont rapportés grâce à deux statistiques : le biais d'estimation et le carré moyen résiduel (CMR) en se basant sur les vrais paramètres de sujet. Les temps de calcul des paramètres de sujet ne seront donc pas rapportés car l'analyse porte uniquement sur la comparaison entre BILOG-MG, ICL et ltm.

Étude 2 : Analyse du TCALS-II (1998)

Dans la deuxième étude, les auteurs ont analysé les données du test de classement en anglais, langue seconde, au collégial (TCALS-II) obtenues au Collège de l'Outaouais en 1998 (Raîche, 2002). Cette épreuve comporte 85 items à choix multiples. Les items 1 à 33 comportent des questions qui mesurent la compréhension auditive. Les items 34 à 70 portent plutôt sur la compréhension à l'écrit. Enfin, les items 71 à 85 sont constitués de questions liées à la lecture de courts textes.

Au total, 1 373 étudiants du cégep de l'Outaouais (749 femmes et 624 hommes) ont répondu à cette épreuve. Il est à noter que Raîche (2002) a déjà démontré que cette épreuve présentait de bonnes qualités métriques : le coefficient alpha de Cronbach global est égal à 0,96 et des analyses factorielles ont démontré que l'unidimensionnalité du trait latent peut être considérée comme une hypothèse valide.

Les résultats découlant de l'analyse du TCALS-II sont étudiés de différentes façons. Premièrement, les auteurs font ressortir certaines statistiques descriptives pour permettre d'entamer la comparaison entre les valeurs du paramètre de difficulté et de sujet (par maximum de vraisemblance et maximum de vraisemblance pondérée) estimées à l'aide des trois logiciels à l'étude. Deuxièmement, ils calculent les coefficients de corrélation de Pearson

entre ces mêmes paramètres. Cela permettra de mettre en évidence la force du lien statistique existant entre les estimations obtenues à l'aide de ces logiciels. Enfin, les temps de calcul seront également rapportés.

Résultats

Étude 1 : résultats pour la simulation

En premier lieu, il est possible d'examiner RCMR. Ces valeurs sont reprises dans le tableau 1. Il est intéressant de noter que ces valeurs ne varient qu'en fonction de la taille d'échantillon : plus celle-ci augmente, plus les RCMR diminuent. C'est un résultat attendu, car l'augmentation de l'information disponible entraîne une augmentation de la précision d'estimation et donc une diminution de la variabilité des estimateurs autour de leurs valeurs réelles. Par contre, les RCMR ne dépendent ni de la longueur du test, ce qui était également prévisible (étant donné que l'estimation se fait item par item), ni de la distribution de probabilité du niveau d'habileté, ce qui était moins prévisible. Enfin, le peu de différences entre les méthodes suggère que les trois logiciels (ltm, ICL et BILOG-MG) produisent des estimations comparables.

Les corrélations moyennes entre les paramètres d'item estimés et réels sont regroupées dans le tableau 2. Il est à constater que ces corrélations ne varient ni en fonction de la méthode, ni en fonction de la longueur du test ou de la distribution de probabilité du niveau d'habileté. Au regard de ce qui précède, c'était là un résultat attendu. De plus, la corrélation augmente lorsque la taille d'échantillon augmente. Ce résultat est en accord avec la tendance des RCMR, car une augmentation de l'information disponible mène à une estimation de meilleure qualité, et de fait à un meilleur ajustement des paramètres estimés aux valeurs réelles. Cela se traduit naturellement par une augmentation de la corrélation. Il est à noter que les corrélations sont toutes supérieures à 0,97, ce qui est en soi un excellent résultat, même pour de petits échantillons.

Tableau 1

RCMR des paramètres d'item estimés, par méthode d'estimation et en fonction de la longueur du test, de la taille du groupe d'étudiants et de la distribution de probabilité du niveau d'habileté (données simulées)

		Distribution normale			*Distribution uniforme*			*Distribution bêta*		
Items	*N*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*
20	100	0,239	0,244	0,244	0,229	0,230	0,230	0,233	0,241	0,240
20	500	0,103	0,106	0,105	0,103	0,104	0,104	0,104	0,107	0,107
20	1000	0,074	0,075	0,075	0,073	0,079	0,079	0,073	0,076	0,076
20	5000	0,032	0,034	0,033	0,034	0,044	0,044	0,033	0,037	0,036
40	100	0,240	0,244	0,244	0,238	0,237	0,237	0,235	0,238	0,238
40	500	0,105	0,107	0,107	0,107	0,108	0,108	0,106	0,108	0,108
40	1000	0,075	0,076	0,076	0,073	0,075	0,074	0,074	0,076	0,076
40	5000	0,033	0,034	0,033	0,034	0,039	0,039	0,034	0,035	0,035
60	100	0,240	0,243	0,246	0,246	0,245	0,246	0,241	0,244	0,247
60	500	0,106	0,108	0,109	0,109	0,109	0,111	0,108	0,109	0,110
60	1000	0,074	0,076	0,076	0,077	0,078	0,078	0,075	0,076	0,077
60	5000	0,034	0,035	0,035	0,035	0,037	0,037	0,034	0,034	0,035
80	100	0,239	0,242	0,255	0,246	0,245	0,255	0,241	0,243	0,257
80	500	0,105	0,106	0,114	0,108	0,108	0,115	0,105	0,107	0,113
80	1000	0,075	0,076	0,080	0,077	0,078	0,082	0,074	0,075	0,080
80	5000	0,034	0,035	0,036	0,034	0,035	0,037	0,033	0,034	0,036

Tableau 2
Corrélations moyennes entre les paramètres d'item estimés et réels, par méthode d'estimation et en fonction de la longueur du test, de la taille du groupe d'étudiants et de la distribution de probabilité du niveau d'habileté (données simulées)

		Distribution normale			*Distribution uniforme*			*Distribution bêta*		
Items	*N*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*
20	100	0,971	0,971	0,971	0,972	0,972	0,972	0,972	0,972	0,972
20	500	0,994	0,994	0,994	0,994	0,994	0,994	0,994	0,994	0,994
20	1000	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997
20	5000	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
40	100	0,972	0,972	0,972	0,972	0,972	0,972	0,972	0,972	0,972
40	500	0,994	0,994	0,994	0,994	0,994	0,994	0,994	0,994	0,994
40	1000	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997
40	5000	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
60	100	0,973	0,973	0,973	0,970	0,970	0,970	0,972	0,972	0,972
60	500	0,994	0,994	0,994	0,994	0,994	0,994	0,994	0,994	0,994
60	1000	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997
60	5000	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
80	100	0,973	0,973	0,973	0,971	0,971	0,971	0,972	0,972	0,972
80	500	0,995	0,995	0,995	0,994	0,994	0,994	0,994	0,994	0,994
80	1000	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997
80	5000	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999

Les auteurs se penchent maintenant sur les temps de calcul requis pour obtenir l'estimation des paramètres d'item. Ils sont présentés dans le tableau 3. En toute logique, il est possible d'observer une augmentation du temps de calcul lorsque la taille de l'échantillon et la longueur du test augmentent : cela est simplement dû au fait que le nombre d'opérations à effectuer est plus important. Cependant, il est intéressant de remarquer que cette augmentation est beaucoup plus marquée avec ltm qu'avec BILOG-MG et ICL. En fait, avec ltm, il est possible d'atteindre des temps de calcul très importants en comparaison avec les deux autres logiciels qui restent très rapides en toute circonstance (jamais plus de quatre secondes de temps de calcul, même pour des échantillons de 5 000 étudiants, en comparaison avec près de cinq minutes pour ltm). De plus, il est intéressant de constater que, quelle que soit la

situation envisagée, ICL est légèrement plus rapide que BILOG-MG, qui est lui-même plus rapide que ltm. Par contre, la distribution de probabilité du niveau d'habileté n'affecte pas le temps de calcul.

Tableau 3

Temps moyens de calcul (en secondes) pour l'estimation des paramètres d'item, par méthode d'estimation et en fonction de la longueur du test, de la taille du groupe d'étudiants et de la distribution de probabilité du niveau d'habileté

		Distribution normale			*Distribution uniforme*			*Distribution bêta*		
Items	*N*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*
20	100	0,90	0,44	0,84	0,87	0,36	0,75	0,84	0,40	0,76
20	500	0,86	0,40	2,84	1,13	0,51	2,87	0,91	0,42	2,86
20	1000	1,11	0,55	5,94	1,32	0,51	6,00	1,12	0,55	6,17
20	5000	2,11	1,59	31,80	2,02	1,26	30,93	2,02	1,63	32,12
40	100	0,92	0,33	2,08	0,91	0,33	2,15	0,94	0,36	2,55
40	500	1,16	0,52	10,12	1,34	0,56	9,29	1,09	0,49	7,94
40	1000	1,33	0,65	16,36	1,32	0,60	16,12	1,33	0,65	16,80
40	5000	2,81	1,98	94,56	2,66	1,91	93,78	2,80	2,01	98,08
60	100	1,04	0,37	3,93	1,04	0,35	3,91	1,04	0,36	3,95
60	500	1,26	0,57	15,85	1,25	0,58	15,82	1,25	0,56	16,05
60	1000	1,47	0,74	33,68	1,47	0,77	33,55	1,46	0,74	34,63
60	5000	3,21	2,51	182,35	3,25	2,62	180,67	3,23	2,50	192,75
80	100	1,16	0,39	6,53	1,15	0,37	6,51	1,18	0,38	6,55
80	500	1,39	0,58	26,44	1,39	0,61	26,32	1,37	0,58	26,63
80	1000	1,65	0,88	53,24	1,65	0,91	53,08	1,63	0,88	54,73
80	5000	3,81	3,12	282,59	3,91	3,26	281,40	3,81	3,13	298,48

Ensuite, les auteurs ont analysé les biais du paramètre de sujet estimé à l'aide des trois logiciels à l'étude. Seront d'abord présentés les résultats pour l'estimateur par maximum de vraisemblance. Comme espéré, il est possible de remarquer que les biais s'amenuisent à mesure que le nombre d'items et que la taille de l'échantillon d'étudiants simulé augmentent. Il est à noter, par contre, que la distribution a un effet sur le biais avec les plus petits échantillons simulés. Plus précisément, il est à constater à la lecture du tableau 4 que pour un échantillon de 20 items et de 100 sujets, le biais est plus élevé

pour des données simulées à partir d'une distribution normale que d'une distribution uniforme, et plus élevé pour des données simulées à partir d'une distribution uniforme que d'une distribution bêta.

Tableau 4
Biais des paramètres de sujet par maximum de vraisemblance et en fonction de la longueur du test, de la taille du groupe d'étudiants et de la distribution de probabilité du niveau d'habileté

		Distribution normale			*Distribution uniforme*			*Distribution bêta*		
Items	*N*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*
20	100	0,012	0,012	0,012	0,007	0,007	0,007	-0,005	-0,005	-0,005
20	500	-0,001	-0,001	-0,001	0,001	0,001	0,001	0,001	0,001	0,001
20	1000	-0,001	-0,001	-0,001	0,000	0,000	0,000	0,000	-0,001	-0,001
20	5000	-0,001	-0,001	-0,001	0,000	0,000	0,000	-0,001	-0,001	-0,001
40	100	-0,005	-0,005	-0,005	0,002	0,002	0,002	0,001	0,001	0,001
40	500	-0,001	-0,001	-0,001	0,000	0,000	0,000	0,000	0,000	0,000
40	1000	0,002	0,002	0,002	-0,001	-0,001	-0,001	0,002	0,002	0,002
40	5000	0,000	0,000	0,000	-0,001	-0,001	-0,001	0,001	0,001	0,001
60	100	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	0,001	0,001	0,001
60	500	0,001	0,001	0,001	-0,001	-0,001	-0,001	0,001	0,001	0,001
60	1000	0,001	0,001	0,001	0,001	0,001	0,001	0,003	0,003	0,003
60	5000	0,000	0,000	0,000	-0,001	-0,001	-0,001	0,001	0,001	0,001
80	100	0,001	0,001	0,001	-0,005	-0,005	-0,005	0,003	0,003	0,003
80	500	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	0,001	0,001	0,001
80	1000	0,001	0,001	0,001	0,000	0,000	0,000	0,001	0,001	0,001
80	5000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

L'étude des RCMR du paramètre de sujet estimé par maximum de vraisemblance permet aussi de démontrer de grandes similitudes pour des échantillons simulés de 40 items et plus. En effet, les résultats présentés au tableau 5 démontrent explicitement le fait que BILOG-MG, ICL et la librairie ltm peuvent être utilisés de façon interchangeable. Par contre, les RCMR simulées à partir de données issues d'une distribution uniforme sont légèrement plus élevées que pour les deux autres distributions.

Tableau 5
RCMR du paramètre de sujet par maximum de vraisemblance et en fonction de la longueur du test, de la taille du groupe d'étudiants et de la distribution de probabilité du niveau d'habileté

		Distribution normale			*Distribution uniforme*			*Distribution bêta*		
Items	*N*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*
20	100	0,577	0,576	0,576	0,613	0,606	0,606	0,569	0,568	0,568
20	500	0,572	0,572	0,572	0,600	0,595	0,595	0,569	0,569	0,569
20	1000	0,575	0,575	0,575	0,598	0,593	0,593	0,571	0,571	0,571
20	5000	0,573	0,573	0,573	0,598	0,593	0,593	0,571	0,571	0,571
40	100	0,389	0,389	0,389	0,408	0,405	0,405	0,400	0,400	0,400
40	500	0,395	0,395	0,395	0,405	0,403	0,403	0,394	0,394	0,394
40	1000	0,395	0,396	0,396	0,408	0,406	0,406	0,393	0,393	0,393
40	5000	0,393	0,393	0,393	0,405	0,403	0,403	0,392	0,392	0,392
60	100	0,320	0,320	0,320	0,326	0,325	0,325	0,321	0,321	0,321
60	500	0,316	0,316	0,316	0,327	0,326	0,326	0,318	0,318	0,318
60	1000	0,319	0,319	0,319	0,326	0,324	0,324	0,315	0,315	0,315
60	5000	0,316	0,317	0,317	0,325	0,324	0,324	0,316	0,316	0,316
80	100	0,274	0,274	0,274	0,280	0,279	0,279	0,275	0,275	0,275
80	500	0,274	0,274	0,274	0,281	0,280	0,280	0,274	0,274	0,274
80	1000	0,274	0,274	0,274	0,281	0,280	0,280	0,273	0,273	0,273
80	5000	0,273	0,273	0,273	0,280	0,279	0,279	0,272	0,272	0,273

Ensuite, les résultats pour l'estimateur par maximum de vraisemblance pondérée seront présentés (voir tableau 6). Comme pour les résultats présentés au tableau 4, il est à noter que les biais s'amenuisent à mesure que le nombre d'items et que le la taille de l'échantillon d'étudiants simulés augmentent. Par contre, pour un échantillon de 20 items et 100 sujets, il apparaît que le biais est plus élevé pour les données simulées à partir d'une distribution normale. Encore une fois, ces résultats prévisibles permettent de mettre en exergue le fait que les trois logiciels à l'étude conduisent à des résultats similaires.

Tableau 6

Biais des paramètres de sujet par maximum de vraisemblance pondérée et en fonction de la longueur du test, de la taille du groupe d'étudiants et de la distribution de probabilité du niveau d'habileté

		Distribution normale			*Distribution uniforme*			*Distribution bêta*		
Items	*N*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*
20	100	0,012	0,012	0,012	0,006	0,006	0,006	-0,006	-0,006	-0,006
20	500	-0,001	-0,001	-0,001	0,001	0,001	0,001	-0,001	-0,001	-0,001
20	1 000	0,000	0,000	0,000	0,000	0,000	0,000	-0,002	-0,002	-0,002
20	5 000	-0,001	-0,001	-0,001	0,000	0,000	0,000	-0,002	-0,002	-0,002
40	100	-0,004	-0,004	-0,004	0,002	0,001	0,001	-0,001	-0,001	-0,001
40	500	-0,001	-0,001	-0,001	0,000	0,000	0,000	-0,002	-0,002	-0,002
40	1 000	0,002	0,002	0,002	-0,001	-0,001	-0,001	0,000	0,000	0,000
40	5 000	0,000	0,000	0,000	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001
60	100	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	0,000	0,000	0,000
60	500	0,001	0,001	0,001	-0,001	-0,001	-0,001	0,000	0,000	0,00
60	1 000	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001
60	5 000	0,000	0,000	0,000	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001
80	100	0,001	0,001	0,001	-0,005	-0,005	-0,005	0,002	0,002	0,002
80	500	-0,001	-0,001	-0,001	-0,001	-0,001	-0,001	0,000	0,000	0,000
80	1 000	0,001	0,001	0,001	0,000	0,000	0,000	0,001	0,001	0,001
80	5 000	0,000	0,000	0,000	0,000	0,000	0,000	-0,001	-0,001	-0,001

Le tableau 7 présente les RCMR du paramètre de sujet par maximum de vraisemblance pondérée. Il est possible de retrouver sensiblement les mêmes conclusions que pour le maximum de vraisemblance : les logiciels à l'étude présentent des résultats comparables. Néanmoins, les auteurs doivent nuancer leur propos en soulignant que la distribution uniforme présente des RCMR légèrement plus élevées que la distribution normale et la distribution bêta.

Tableau 7
RCMR des paramètres de sujet par maximum de vraisemblance pondérée et en fonction de la longueur du test, de la taille du groupe d'étudiants et de la distribution de probabilité du niveau d'habileté

		Distribution normale			*Distribution uniforme*			*Distribution bêta*		
Items	*N*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*	*BILOG-MG*	*ICL*	*ltm*
20	100	0,539	0,539	0,539	0,563	0,558	0,558	0,534	0,533	0,533
20	500	0,536	0,535	0,535	0,554	0,550	0,550	0,533	0,533	0,533
20	1 000	0,538	0,538	0,538	0,552	0,548	0,548	0,535	0,535	0,535
20	5 000	0,536	0,536	0,536	0,552	0,548	0,548	0,535	0,535	0,535
40	100	0,374	0,374	0,374	0,391	0,389	0,389	0,385	0,385	0,385
40	500	0,380	0,380	0,380	0,389	0,387	0,387	0,380	0,380	0,380
40	1 000	0,080	0,080	0,080	0,391	0,389	0,389	0,378	0,379	0,379
40	5 000	0,379	0,379	0,379	0,389	0,387	0,387	0,378	0,378	0,378
60	100	0,312	0,312	0,312	0,317	0,316	0,316	0,314	0,314	0,314
60	500	0,308	0,309	0,309	0,318	0,317	0,317	0,310	0,311	0,311
60	1 000	0,311	0,311	0,311	0,317	0,316	0,316	0,308	0,308	0,308
60	5 000	0,309	0,309	0,309	0,317	0,316	0,316	0,309	0,309	0,309
80	100	0,269	0,269	0,269	0,274	0,273	0,273	0,270	0,270	0,270
80	500	0,269	0,269	0,269	0,276	0,275	0,275	0,269	0,269	0,269
80	1 000	0,269	0,269	0,269	0,276	0,275	0,275	0,268	0,268	0,268
80	5 000	0,268	0,268	0,268	0,274	0,274	0,274	0,268	0,268	0,268

Les résultats présentés aux tableaux précédents permettent de montrer que les logiciels présentent des biais d'estimation et des RCMR comparables, et ce, peu importe qu'il s'agisse de BILOG-MG, ICL ou de la librairie ltm.

Étude 2 : Résultats pour le TCALS-II (1998)

Par suite de la mise à l´échelle des estimations par maximum de vraisemblance marginale, des moyennes et des écarts types de même amplitude ont été obtenus pour les trois logiciels étudiés. Le tableau 8 rapporte d´ailleurs des valeurs moyennes, des écarts types et des étendues fort comparables. Comme prévu, les valeurs moyennes sont nulles à la suite de la calibration des paramètres des items et les écarts types sont proches de l'unité. Il y a peu

de variations dans les quartiles d'une méthode à l'autre, excepté pour ICL qui présente une dispersion des paramètres légèrement inférieure aux deux autres méthodes.

Tableau 8
Statistiques descriptives des paramètres de difficulté estimés

	Min	*1er quartile*	*Médiane*	*Moyenne*	*3e quartile*	*Max*	*Écart type*
BILOG-MG	-2,44	-0,63	-0,18	0,00	0,75	2,93	1,09
ICL	-2,40	-0,61	-0,17	0,00	0,73	2,81	1,06
ltm	-2,44	-0,64	-0,19	0,00	0,73	2,89	1,09

Les corrélations de Pearson indiquent aussi un aperçu convaincant de la similitude des estimations obtenues par BILOG-MG, ICL et ltm. Comme constaté au tableau 9, ces paramètres sont tous fortement corrélés.

Tableau 9
Corrélations de Pearson pour le paramètre de difficulté estimé par maximum de vraisemblance marginale

	BILOG-MG	*ICL*	*ltm*
BILOG-MG	1,00		
ICL	0,99	1,00	
ltm	0,99	0,99	1,00

L'ordre de grandeur des temps de calcul présenté à la section précédente est le même. La librairie ltm a pris 54,03 secondes avant d'obtenir des résultats comparativement à 0,92 seconde pour ICL et à 1,69 seconde pour BILOG-MG.

De leur côté, les paramètres d'habileté présentent aussi une grande similitude, et cela, peu importe la méthode d'estimation ou le logiciel qui a été utilisé. Par exemple, le tableau 10 présente les moyennes et les écarts types similaires pour l'analyse du TCLAS-II.

Tableau 10
Statistiques descriptives des paramètres de sujet estimés

	Min	*1er quartile*	*Médiane*	*Moyenne*	*3e quartile*	*Max*	*Écart type*
MV BILOG-MG	-1,78	0,80	2,08	1,92	3,08	4,00	1,47
MV ICL	-1,76	0,80	2,06	1,90	3,05	4,00	1,46
MV ltm	-1,76	0,79	2,06	1,90	3,05	4,00	1,46
MVP BILOG-MG	-1,76	0,79	2,05	1,89	3,01	4,00	1,45
MVP ICL	-1,74	0,78	2,03	1,87	2,98	4,00	1,44
MVP ltm	-1,74	0,78	2,03	1,87	2,98	4,00	1,44

MV : maximum de vraisemblance ; MVP : maximum de vraisemblance pondérée

De plus, la lecture du tableau 11 est explicite : peu importe l'approche utilisée, les paramètres de sujet sont fortement corrélés.

Tableau 11
Corrélations de Pearson pour le paramètre de sujet estimé

	MV BILOG-MG	*MV ICL*	*MVP MV ltm*	*MVP BILOG-MG*	*MVP ICL*	*ltm*
MV BILOG-MG	1,00					
MV ICL	0,99	1,00				
MV ltm	0,99	1,00	1,00			
MVP BILOG-MG	0,99	0,99	0,99	1,00		
MVP ICL	0,99	0,99	0,99	0,99	1,00	
MVP ltm	0,99	0,99	0,99	0,99	1,00	1,00

MV : maximum de vraisemblance ; MVP : maximum de vraisemblance pondérée

Les résultats obtenus à l'aide de l'estimateur par maximum de vraisemblance et l'estimateur par vraisemblance pondérée offrent donc des résultats comparables.

Discussion et conclusion

Cet article avait comme objectif spécifique de comparer l'estimation des paramètres d'item et de sujet selon une des modélisations issues de la TRI : le modèle de Rasch. Ici, trois logiciels ont été sélectionés : BILOG-MG, ICL et la librairie ltm disponible pour le logiciel R. Ce choix est basé sur des critères qui ont déjà été discutés préalablement mais, citons le fait que ces logiciels présentent certaines différences d'intérêt (par ex., BILOG-MG est payant alors que ltm est gratuit) et qu'ils utilisent une méthode d'estimation des items commune.

Les résultats ont permis de conclure que ces logiciels permettent d'obtenir des résultats similaires, voire interchangeables, pour l'analyse de données selon le modèle de Rasch. Cette similitude est confirmée par l'étude de données simulées (étude 1) qui a mis en évidence des RCMR très proches pour les trois méthodes. L'analyse des données du TCALS-II (étude 2) montre aussi une grande ressemblance entre les paramètres de difficulté estimés.

Étant donné que les résultats des trois logiciels comparés sont similaires, il convient de les distinguer selon un autre critère : le temps d'exécution des estimations. Le tableau 3 et les résultats présentés à l'étude 2 ont, en effet, clairement mis en évidence que les méthodes d'estimation utilisées par ltm sont nettement plus lentes que les méthodes BILOG-MG et ICL. Ceci s'explique notamment par le fait que ltm est une fonction définie « en interne » dans R tandis que BILOG-MG et ICL sont des programmes indépendants programmés en language C, pouvant être appelés depuis R, mais procédant à des calculs de façon plus rapide. De plus, étant donné que ICL est légèrement plus rapide que BILOG-MG et qu'il est disponible gratuitement (alors que BILOG-MG est un logiciel commercial), les auteurs recommandent l'utilisation de ICL pour la calibration selon le modèle de Rasch lorsque le nombre de répondants est très grand.

Les analyses produites dans cet article mènent à recommander l'utilisation de logiciels gratuits pour procéder à l'étude de données à l'aide du modèle de Rasch. En effet, les avantages des logiciels gratuits R et ICL sont doubles. Premièrement, ces logiciels permettent d'utiliser de nombreux modèles de la TRI : autant pour analyser des données à réponses dichotomiques que des données à réponses polytomiques. BILOG-MG, de son côté, ne permet pas de produire des analyses aussi diversifiées. Deuxièmement, R et ICL sont disponibles à un maximum d'utilisateurs grâce à leur gratuité. À l'opposé, BILOG- MG est relativement dispendieux.

Les logiciels R et ICL comportent aussi quelques désavantages. Par exemple, ils imposent une connaissance de base des règles de programmation. Deuxièmement, il est à noter que les librairies disponibles dans R ne sont pas systématiquement évaluées, ce qui exige de l'utilisateur une certaine prudence dans l'utilisation des résultats (notamment lorsqu'ils sont issus de librairies récentes). Enfin, tout n'est pas encore disponible dans les logiciels R et ICL : de nombreux développement liés à la TRI sont encore à produire.

Cette étude présente quelques limites qu'il est important de soulever pour mettre les résultats en contexte. Par exemple, les comparaisons ont été produites entre trois logiciels alors qu'il en existe d'autres qui sont aussi très pertinents : RUMM2030, Xcalibre, PARSCALE ou PARAM-3PL. De plus, il aurait été intéressant de procéder à des comparaisons basées sur des données réelles provenant de tests de tailles différentes. Dans le cadre de cette étude, l'analyse est restreinte à une seule matrice de données : le test de classement en anglais, langue seconde, au collégial (TCALS-II).

Plusieurs études doivent être entreprises afin d'améliorer la compréhension et l'utilisation de la TRI sous d'autres logiciels gratuits. Premièrement, des recherches doivent être conduites afin de valider les résultats des paramètres estimés à partir des autres modèles de la TRI. Deuxièmement, il serait intéressant de développer des interfaces facilitant leur utilisation. Pour rappel, les chercheurs en éducation apprécient surtout SPSS pour sa facilité d'utilisation ; il serait pertinent de s'en inspirer pour faciliter l'utilisation de R. Troisièmement, il reste encore beaucoup de modules supplémentaires à développer pour appliquer la TRI avec les logiciels gratuits. Par exemple, outre plink, il existe peu de librairies pour produire des analyses d'appariement vertical et horizontal (*equating*).

NOTE

1. Le web 2.0 est l'évolution du web vers une forme plus conviviale et plus simple. Par exemple, cette version intègre l'utilisation de logiciels gratuits et propose des plateformes qui facilitent l'échange entre les utilisateurs. Le lecteur intéressé à en savoir plus pourra consulter Alexander (2006).

RÉFÉRENCES

Alexander, B. (2006). Web 2.0: A new wave of innovation for teaching and learning? *EDUCAUSE Review, 41,* 32–44.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2e éd.). New York, NY: Dekker.

Benzécri, J.-P. (1973). La place de l'a priori. *Encyclopaedia universalis, Organum, 17*, 11-24.

Bertrand, R., & Blais, J.-G. (2004). *Modèles de mesure : L'apport de la théorie de la réponse aux items*. Sainte-Foy, Canada : Presses de l'Université du Québec.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison–Wesley.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1-28.

Forsyth, R., Sarsangjan, V., & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement, 5*, 175-186. doi: 10.1177/014662168100500203

Hambleton, R. K., & Swaminathan, H. (1985). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hanson, B. A. (2002). *IRT Command Language (ICL). Computer software version 0.020301*. Retrieved from http://www.b-a-h.com/software/irt/icl/index.html

Hulin, C.L., Drasgow, F., & Parson, C.K. (1983). *Item response theory- Application to psychological measurement*. Homewood, IL: Irwin.

Jurich, D., & Goodman, J.T. (2009, October). *Comparison IRT parameter recovery of mixed format examinations in PARSCALE and ICL*. Poster session presented at the meeting of Northeastern Educational Research Association, Rocky Hill, Connecticut.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking (2nd edition)*: *Methods and practice*. New York, NY: Springer.

Laurier, M., Froio, L., Pearo, C., & Fournier, M. (1998). *Test de classement d'anglais langue seconde au collégial. Rapport technique*. Document inédit, Collège de Maisonneuve, Montréal, Canada.

Lord, F. M. (1952). *A Theory of Test Scores* (Psychometric Monograph No. 7). Richmond, VA: Psychometric Corporation.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York, NY: Lawrence Erlbaum.

Magis, D., Béland, S., & Raîche, G. (2012). *difR: Collection of methods to detect dichotomous differential item functioning (DIF) in psychometrics. R package version 4.2*. Retrieved from http://CRAN.R-project.org/package=difR

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior research methods*, *42*, 847-862. doi: http://dx.doi.org/10.3758/BRM.42.3.847

Magis, D., & Raîche, G. (2011). catR: an R package for computerized adaptive testing. *Applied Psychological Measurement*, *35*, 576-577. doi: http://dx.doi.org/10.1177/0146621611407482

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8), 1-31.

Mair, P., & Hatzinger, R. (2007a). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*, 1-20.

Mair, P., & Hatzinger, R. (2007b). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science, 49*, 26-43.

Mair, P., Hatzinger, R., & Maier, M. (2010). *eRm: Extended Rasch Modeling. R package version 0.13-0*. Retrieved from http://CRAN.R-project.org/package=eRm

Mead, A. D., Morris, S. B., & Blitz, D. L. (2007). *Open-source IRT: A comparison of BILOG-MG and ICL features and item parameter recovery.* Unpublished document. Retrieved from URL: http://mypages.iit.edu/~mead/MeadMorrisBlitz2007.pdf

Partchev, I. (2011). *Irtoys: Simple interface to the estimation and plotting of IRT models. R package version 0.1.4*. Retrieved from http://CRAN.R-project.org/package=irtoys

R Development Core Team (2012). *R: A language and environment for statistical computing*. Vienne, Autriche: R Foundation for Statistical Computing.

Raîche, G. (2002). *Le dépistage du sous-classement aux tests de classement en anglais, langue seconde, au collégial*. Document inédit, Gatineau, Canada: Collège de l'Outaouais.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: The University of Chicago Press.

Rentz, R. R., & Barshaw, W. L. (1977). The national reference scale for reading: An application of the Rasch model. *Journal of Educational Measurement, 14*, 161-180.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*, 1–25.

Rizopoulos, D. (2012). *ltm. R package version 0.9-7*. Retrieved from http://CRAN.R-project.org/ package=ltm

Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing, 3*, 365-384. doi: http://dx.doi.org/10.1207/S15327574IJT0304_5

Sijtsma, K., & Junker B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika, 33,* 75-102. doi: 10.2333/ bhmk. 33.75

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika*, *54*, 427-450.

Weeks, J. P. (2010). Plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software*, *35*, 1-33.

Weeks, J. P. (2011). *plink. R package version 1.3-1*. Retrieved from http://CRAN.R-project.org/ package=plink

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

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]*. Chicago, IL: Scientific Software International.

Date de réception : 27 juin 2012

Date de réception de la version finale : 21 décembre 2012

Date d'acceptation : 27 décembre 2012