Journal of Educational and Behavioral Statistics
2015, Vol. 40, No. 1, pp. 96–105
DOI: 10.3102/1076998614558122
© 2014 AERA. http://jebs.aera.net

A Note on the Equivalence Between Observed and Expected Information Functions With Polytomous IRT Models

David Magis
KU Leuven and University of Liège, Belgium

The purpose of this note is to study the equivalence of observed and expected (Fisher) information functions with polytomous item response theory (IRT) models. It is established that observed and expected information functions are equivalent for the class of divide-by-total models (including partial credit, generalized partial credit, rating scale, and nominal response models) but not for the class of difference models (including the graded response and modified graded response models). Yet, observed information function remains positive in both classes. Straightforward connections with dichotomous IRT models and further implications are outlined.

Keywords: *item response theory; polytomous models; observed information function; expected information function*

Introduction

In item response theory (IRT), information functions are useful tools to evaluate the contribution of an item to some testing procedure or for ability estimation purposes. Two types of information functions exist: the observed item information, defined as the curvature (i.e., minus the second derivative) of the log-likelihood function of ability and the expected (or Fisher) item information, which is the mathematical expectation of the observed information function. The observed information directly depends on the particular item responses provided by the test takers, while expected information does not.

In the context of dichotomous item responses, the discrepancy between observed and Fisher information functions was investigated by Bradlow (1996), Samejima (1973), van der Linden (1998), Veerkamp (1996), and Yen, Burket, and Sykes (1991), among others. All findings can be summarized as follows. First, under the two-parameter logistic (2PL) model, and subsequently under the Rasch model, both types of information functions are completely equivalent. Second, under the three-parameter logistic (3PL) model, observed

Downloaded from http://jebs.aera.net at KU Leuven University Library on March 7, 2016

and Fisher information are usually different. Third, Samejima (1973) and Yen et al. highlighted that the observed information function can take negative values and discussed practical situations wherein negative information can become problematic (see also Bradlow, 1996).

Although straightforward extensions to polytomously scored items were mentioned (Bradlow, 1996), it seems that such a comparison has not been performed so far. Consequently, the purpose of this note is to describe the relationships between observed and expected information functions with polytomous IRT models.

Polytomous IRT Models and Information Functions

Consider a test of n items and for each item $j(j = 1, \ldots, n)$, let $g_j + 1$ be the number of response categories. Moreover, X_j is the item score, with $X_j \in \{0, 1, \ldots, g_j\}$, and $\boldsymbol{\xi}_j$ is the vector of item parameters (depending on the model and the number of item scores). Finally, $P_{jk}(\theta) = \Pr(X_j = k|\theta, \boldsymbol{\xi}_j)$ is the probability of score $k(k = 0, \ldots, g_j)$ for a test taker with ability level θ. Two classes of polytomous IRT models are considered: the so-called difference models and divide-by-total models (Thissen & Steinberg, 1986).

Difference and Divide-By-Total Models

Difference models are defined by means of cumulative probabilities $P^*_{jk}(\theta) = \Pr(X_j \geq k|\theta, \boldsymbol{\xi}_j)$ (i.e., the probability of score k, $k + 1, \ldots,$ or g_j) with $P^*_{j0}(\theta) = 1$, $P^*_{j,g_j}(\theta) = P_{j,g_j}(\theta)$ and the convention that $P^*_{jk}(\theta) = 0$ for any $k > g_j$. Score probabilities are found as $P_{jk}(\theta) = P^*_{jk}(\theta) - P^*_{j,k+1}(\theta)$. The cumulative probabilities take the following form:

$$P^*_{jk}(\theta) = \frac{e_{jk}(\theta)}{1 + e_{jk}(\theta)}, \tag{1}$$

where $e_{jk}(\theta)$ is a positive increasing function of θ and such that its first derivative $e'_{jk}(\theta)$ (with respect to θ) takes the following form:

$$e'_{jk}(\theta) = \frac{de_{jk}(\theta)}{d\theta} = \alpha_j e_{jk}(\theta), \tag{2}$$

α_j is strictly positive and independent of both k and θ. This class encompasses the graded response model (GRM; Samejima, 1969, 1996) and the modified graded response model (Muraki, 1990).

Divide-by-total models take the following form:

$$P_{jk}(\theta) = \frac{\Gamma_{jk}(\theta)}{\sum_{r=0}^{g_j} \Gamma_{jr}(\theta)} = \frac{\Gamma_{jk}(\theta)}{\Gamma_j(\theta)}, \tag{3}$$

where $\Gamma_{jk}(\theta)$ is a positive function of θ, such that its first derivative (with respect to θ) is given by:

Downloaded from http://jebs.aera.net at KU Leuven University Library on March 7, 2016

$$\Gamma'_{jk}(\theta) = \frac{d\Gamma_{jk}(\theta)}{d\theta} = h_{jk}\Gamma_{jk}(\theta), \tag{4}$$

and h_{jk} does not depend on θ. This class encompasses the partial credit model (PCM; Masters, 1982), the generalized PCM (Muraki, 1992), the rating scale model (Andrich, 1978a, 1978b), and the nominal response model (NRM; Bock, 1972), among others.

Observed and Expected Information Functions

Independently of the class of polytomous IRT model, observed and Fisher information functions can be defined from a global modeling approach. Set first τ_{jk} as a stochastic indicator being equal to 1 if $X_j = k$ and 0 otherwise. The observed item information function (OIF), also sometimes referred to as the item response information function (Samejima, 1969, 1994), is defined as

$$OI_j(\theta) = -\frac{d^2}{d\theta^2}\log L_j\left(\theta|X_j\right), \tag{5}$$

where $L_j\left(\theta|X_j\right)$ is the likelihood function of ability θ for item j with item score X_j:

$$L_j(\theta|X_j) = \prod_{k=0}^{g_j} P_{jk}(\theta)^{\tau_{jk}}. \tag{6}$$

The expected item information function (EIF), or simply item information function (Samejima, 1969, 1994), is defined as the mathematical expectation of the observed information function:

$$I_j(\theta) = E[OI_j(\theta)] = -E\left[\frac{d^2}{d\theta^2}\log L_j(\theta|X_j)\right]. \tag{7}$$

Equations 6 and 7 also hold for simple dichotomous IRT models, for which it was shown that $OI_j(\theta)$ and $I_j(\theta)$ functions are perfectly identical under the 2PL model but not under the 3PL model (Bradlow, 1996; Samejima, 1973; Yen, Burket, & Sykes, 1991). It was also pointed out that the OIF can be negative with the 3PL model, thus jeopardizing its use for, for example, computation of asymptotic standard errors (*SE*s) of ability estimates or next item selection in computerized adaptive tests. The EIF, on the other hand, is always positive even under the 3PL model, which motivates its use for computing the test information function (TIF) $I(\theta) = \sum_{j=1}^{n} I_j(\theta)$ (Birnbaum, 1968).

Equivalence With Polytomous Models

The purpose of this section is to establish the three following results: (a) observed and expected information functions are completely equivalent with divide-by-total models, (b) they are usually different with difference models, (c) whatever the class of polytomous models, the OIF always takes positive values. Only the main mathematical steps are sketched subsequently. A detailed development is provided in Magis (2014) and available upon request.

Downloaded from http://jebs.aera.net at KU Leuven University Library on March 7, 2016

The starting point consists in rewriting the OIF and EIF under the simpler following forms:

$$OI_j(\theta) = \sum_{k=0}^{g_j} \tau_{jk} \left[\frac{P'_{jk}(\theta)^2}{P_{jk}(\theta)^2} - \frac{P''_{jk}(\theta)}{P_{jk}(\theta)} \right] \text{ and } I_j(\theta) = \sum_{k=0}^{g_j} \left[\frac{P'_{jk}(\theta)^2}{P_{jk}(\theta)} - P''_{jk}(\theta) \right], \quad (8)$$

where $P'_{jk}(\theta)$ and $P''_{jk}(\theta)$ stand, respectively, for the first and second derivatives of P_{jk}. These Equations are obtained by explicitly writing down the first and second derivatives of the log-likelihood function $\log L_j(\theta|X_j)$ with respect to θ. Let us investigate now the conditions for which $OI_j(\theta)$ and $I_j(\theta)$ in Equation 8 are identical.

We start with difference models defined by Equations 1 and 2. By explicitly writing the derivatives $P^{*'}_{jk}(\theta)$, $P'_{jk}(\theta)$ and $P''_{jk}(\theta)$, and plugging-in them into Equation 8, it is straightforward to establish that

$$OI_j(\theta) = \sum_{k=0}^{g_j} \tau_{jk} \alpha_j \left[P^{*\prime}_{jk}(\theta) + P^{*\prime}_{j,k+1}(\theta) \right] \quad (9)$$

and

$$I_j(\theta) = \sum_{k=0}^{g_j} P_{jk}(\theta) \alpha_j \left[P^{*\prime}_{jk}(\theta) + P^{*\prime}_{j,k+1}(\theta) \right]. \quad (10)$$

Equations 9 and 10 lead to two important results. First, the OIF will always be positive, whatever the item parameters and the ability level, since the cumulative probabilities $P^*_{jk}(\theta)$ and $P^*_{j,k+1}(\theta)$ are increasing with θ and α_j is strictly positive by definition. Second, despite the closeness of Equations 9 and 10, it is neither obvious nor necessary that both OIF and EIF functions will always coincide. In order to better see the discrepancy between these two kinds of information functions, define $T_{jk}(\theta)$ as

$$T_{jk}(\theta) = \alpha_j^2 \left\{ P^*_{jk}(\theta) \left[1 - P^*_{jk}(\theta) \right] + P^*_{j,k+1}(\theta) \left[1 - P^*_{j,k+1}(\theta) \right] \right\}, \quad (11)$$

so that Equations 9 and 10 can be simply written as:

$$OI_j(\theta) = \sum_{k=0}^{g_j} \tau_{jk} T_{jk}(\theta) \text{ and } I_j(\theta) = \sum_{k=0}^{g_j} P_{jk}(\theta) T_{jk}(\theta). \quad (12)$$

In other words, the OIF takes the value of $T_{jk}(\theta)$ that corresponds to the response category under consideration, while the EIF is the weighted average of all $T_{jk}(\theta)$ values, with the weights equal to the response category probabilities. In very particular situations, OIF and EIF values may coincide with difference models. This is the case, for instance, when the item has only two possible scores (i.e., when $g_j = 1$). This corresponds exactly to the simple 2PL model with dichotomous items, for which the equivalence between OIF and EIF was already established.

We focus now on divide-by-total models defined by Equations 3 and 4. By writing explicitly the first and second derivatives of $P_{jk}(\theta)$, it is easy to establish that

$$\frac{P'_{jk}(\theta)^2}{P_{jk}(\theta)^2} - \frac{P''_{jk}(\theta)}{P_{jk}(\theta)} = \sum_{r=0}^{g_j} h_{jr} P_{jr}(\theta) \left[h_{jr} - \sum_{t=0}^{g_j} h_{jt} P_{jt}(\theta) \right]. \quad (13)$$

Downloaded from http://jebs.aera.net at KU Leuven University Library on March 7, 2016

The right-hand side of Equation 13 does not depend on the particular response category k. Denoting it by C_j, it follows that from Equation 8 that $OI_j(\theta) = \sum_{k=0}^{g_j} \tau_{jk} C_j = C_j$. This implies that the OIF takes the same value, independently of the particular item score. Moreover, it can be shown that

$$\frac{P'_{jk}(\theta)^2}{P_{jk}(\theta)} - P''_{jk}(\theta) = P_{jk}(\theta) C_j, \tag{14}$$

leading to $I_j(\theta) = \sum_{k=0}^{g_j} P_{jk}(\theta) C_j = C_j$. In sum, the OIF and EIF coincide for any item and ability level under divide-by-total models.

Practical Implications

With dichotomous IRT models (and more precisely the 3PL model), Bradlow (1996) pointed out three important practical issues when negative information functions occur: (a) convergence issues of model-fitting algorithms that make use of observed information, such as Newton–Raphson; (b) negative variance estimates when using quadratic approximation of the log-likelihood function and related to the previous drawback, and (c) problems in, for example, Bayesian methods such as Gibbs sampling that make use of a Metropolis–Hastings step with a quadratic approximation of the log-likelihood function. With polytomous IRT models, however, these issues vanish since the OIF will always return positive values, whatever the class of models. This is a clear asset in this context: One does not need to worry about possible aforementioned computational issues, even when considering observed information functions. Moreover, with divide-by-total models, the equivalence between OIF and EIF avoids the discussion about which type of information to select (since both will return the same values). However, with difference models, the gap between observed and expected information functions should be further studied, as it can have some non-negligible impact on related measures.

One example is the computation of the *SE* of ability estimates, which often depends directly on the information function. For instance, the *SE* of the maximum likelihood (ML) estimator of ability is computed as the inverse of the square root of the TIF (Birnbaum, 1968). Thus, the two information functions would lead to different *SE* values for the same response pattern and ability estimate, with probably a larger gap with fewer items (and hence less item contributions to the TIF). This may have a straightforward impact in, for example, computerized adaptive testing (CAT). If item selection rules are often based on the EIF only, it is common practice to use a stopping rule based on the precision of the adinterim ability estimate. Choosing either the OIF or the EIF could lead to non-negligible differences in the values of the *SE* of ability and hence may impact considerably the length of the CAT.

An artificial yet illustrative example is used for this purpose. An artificial item bank of 100 items was generated under the GRM with at most four response categories per item. Then, after picking up randomly the first item to administer, item

Downloaded from http://jebs.aera.net at KU Leuven University Library on March 7, 2016

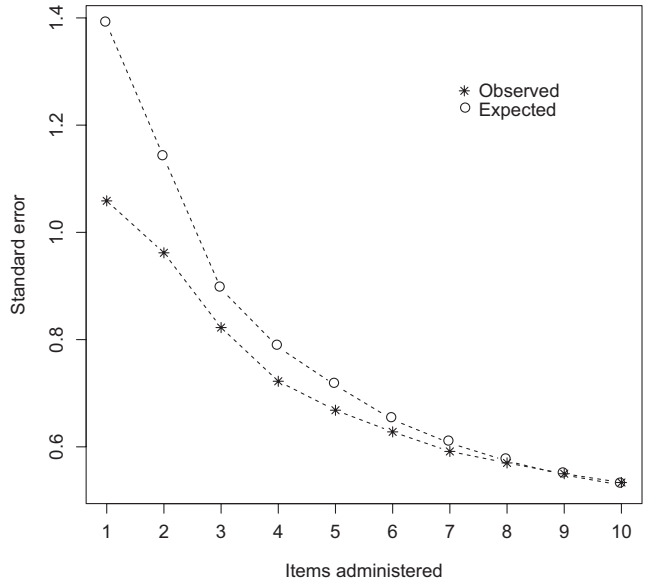

FIGURE 1. *Standard error of ML estimates of ability for one generated CAT response pattern, using observed or expected information function. ML = maximum likelihood; CAT = computerized adaptive testing.*

responses were drawn for a test taker of true ability level of zero, and the adinterim ability estimates were obtained with the usual ML method. The next item was selected using maximum Fisher information (MFI) rule (i.e., the item whose expected information function is maximal for the adinterim ability estimate). The CAT stopped once 10 items were administered. Then, at each stage of the adaptive administration, the *SE* of the adinterim ability estimate was recomputed, using either the EIF or the OIF, leading to two sets of 10 *SE* values (one for each type of information function). The R package *catR* (Magis & Raîche, 2012) of the R environment (R Core Team, 2014) was used for that purpose. The complete R code for this example is available upon request.

Figure 1 displays the *SE* values as a function of the number of items administered in the CAT (from 1 to 10). As the number of administered items increases, the *SE* values (derived with both types of information functions) decrease, which is obvious by definition of the TIF and the aforementioned results. Note also that the *SE* values obtained with the OIF are smaller than or equal to the corresponding values with the EIF. However, this is mostly due to the random generation of response patterns, and this trend is not systematic (as highlighted by other simulated patterns, not shown here). Most important, the gap between the *SE* values decreases quickly with the test length. This highlights that the choice of a particular type of information function, though leading to substantial differences in *SE*

Downloaded from http://jebs.aera.net at KU Leuven University Library on March 7, 2016

TABLE 1
Computation of Observed and Expected Item Information Functions for an Artificial Four-category Item Calibrated Under the Graded Response Model, With Parameters $\alpha_j = 1.2$, $\beta_{j1} = -1.4$, $\beta_{j2} = 0.2$, *and* $\beta_{j3} = 1$, *and for Ability Level* $\theta = 0$

	Response category				
	$k = 0$	$k = 1$	$k = 2$	$k = 3$	Sum
$e_{jk}(\theta)$	NA	5.366	0.787	0.301	NA
$P^*_{jk}(\theta)$	1.000	0.843	0.440	0.231	NA
$P_{jk}(\theta)$	0.157	0.403	0.209	0.231	1.000
$T_{jk}(\theta)$	0.191	0.546	0.611	0.256	NA
$P_{jk}(\theta)T_{jk}(\theta)$	0.030	0.220	0.128	0.059	0.436

Note. NA = no quantity has been computed in this table cell.

values with few items, becomes negligible with longer tests. Though this trend was observed throughout all simulated designs, this has to be further evaluated through systematic comparative studies (involving different difference models, item designs, and CAT options).

Illustration

To end up, the computation of observed and expected information functions is illustrated with a simple artificial example. All steps are described in detail for illustration purposes, but such calculations can be routinely performed with, for example, the R package *catR*. Consider an item with four possible responses, coded from 0 to 3 for, say, responses *Never*, *Sometimes*, *Often*, and *Always*. Assume the item belongs to a broader test of polytomous items that was calibrated under the GRM, for which the functions $e_{jk}(\theta)$ ($k = 1, 2, 3$) are equal to $\exp[\alpha_j(\theta - \beta_{jk})]$ (using the notations by Embretson & Reise, 2000). Set the item parameters as follows: $\alpha_j = 1.2$, $\beta_{j1} = -1.4$, $\beta_{j2} = 0.2$, and $\beta_{j3} = 1$. Using these parameter values, it is straightforward to derive successively (and for any ability level θ): (a) the functions $e_{jk}(\theta)$ ($k = 1, 2, 3$); (b) the cumulative probabilities $P^*_{jk}(\theta)$ ($k = 1, 2, 3$), using Equation 1 and $P^*_{j0}(\theta) = 1$; (c) the response category probabilities $P_{jk}(\theta) = P^*_{jk}(\theta) - P^*_{j,k+1}(\theta)$ ($k = 0, 1, 2, 3$), with $P^*_{j4}(\theta) = 0$.

Table 1 lists all these computation steps for the particular value $\theta = 0$. As pointed out by Equation 12, each value of $T_{jk}(\theta)$ (fourth column of Table 1) actually corresponds to the OIF for each particular response category. Furthermore, the EIF corresponds to the sum of all $T_{jk}(\theta)$ values, weighted by the response category probabilities $P_{jk}(\theta)$ (listed in the fifth column of Table 1). As anticipated, the expected information function (0.436) does not correspond to any value of the observed information functions (at ability level zero).

Downloaded from http://jebs.aera.net at KU Leuven University Library on March 7, 2016

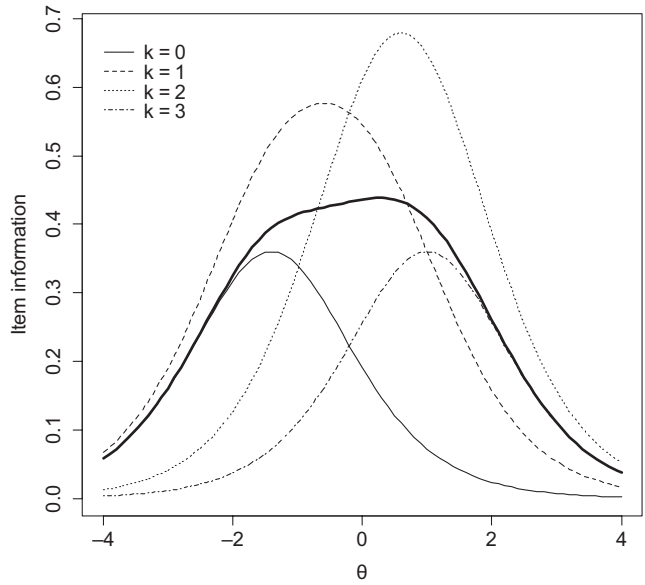

FIGURE 2. *Observed and expected item information functions for the artificial item. Expected information function is displayed with a thick plain trait.*

The observed and expected information functions are also graphically displayed on Figure 2, for a wide range of ability levels. The EIF can be clearly seen as a weighted average of all individual OIFs. At lower ability levels, response category 0 is the most probable and hence the weights favor this score, leading to an EIF that is very close to the OIF for score 0. Similarly, at very large ability levels, the EIF receives almost all weight from the last item score ($k = 3$ in this example) since the probability for the latter is almost equal to 1. In between these extremes, both all individual score probabilities and OIF contribute to the EIF.

Final Comments

The purpose of this article was to extend the comparison of observed EIFs to polytomous IRT models, by setting a general framework first and by considering two broad classes of polytomous IRT models. It was established first that both kinds of functions coincide perfectly for the class of divide-by-total models, while it is not necessarily (and actually rarely) the case for difference models. The conceptual gap between difference and divide-by-total models originates from the fact that the OIF is a weighted sum of terms that depend on the response category for difference models but not for divide-by-total models. This gap is therefore an intrinsic characteristic of the main classes of models. However, the

Downloaded from http://jebs.aera.net at KU Leuven University Library on March 7, 2016

OIF can only take positive values (independently of the choice of the polytomous model), in contrast to what happens to the 3PL model.

Though the distinction between observed and expected information functions is crucial with dichotomous IRT models (and especially the 3PL model), it is of less practical relevance with polytomous models. Technical issues related to negative information functions will never occur in this framework, and both types of information are completely equivalent with divide-by-total models. Thus, the discussion of which information function to select is meaningful only for the class of difference models. One rule of thumb can be to focus on expected information as the by default function (as usually done for dichotomous models) and evaluate the impact of replacing it by the observed function in the framework of interest.

Acknowledgment

The author thanks the editors and anonymous reviewers for insightful comments and suggestions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the Research Fund of KU Leuven, Belgium (GOA/15/003) and the Interuniversity Attraction Poles programme financed by the Belgian government (IAP/P7/06).

References

Andrich, D. (1978a). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*, 581–594. doi: 10.1177/014662167800200413

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. doi:10.1007/BF02293814

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.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51. doi:10.1007/BF02291411

Bradlow, E. T. (1996). Negative information and the three-parameter logistic model. *Journal of Educational and Behavioral Statistics*, *21*, 179–185. doi:10.2307/1165216

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Magis, D. (2014). *On the equivalence between observed and expected information functions with polytomous IRT models*. Research Report, Department of Psychology, KU Leuven, Leuven, Belgium.

Downloaded from http://jebs.aera.net at KU Leuven University Library on March 7, 2016

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi:10.1007/BF02296272

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*, 59–71. doi:10.1177/014662169001400106

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176. doi:10.1177/014662169201600206

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.

Samejima, F. (1973). A comment to Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika*, *38*, 221–233. doi:10.1007/BF02291115

Samejima, F. (1994). Some critical observations of the test information function as a measure of local accuracy in ability estimation. *Psychometrika*, *59*, 307–329. doi:10.1007/BF02296127

Samejima, F. (1996). The graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York, NY: Springer.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577. doi:10.1007/BF02295596

van der Linden, W. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*, 201–216. doi:10.1007/BF02294775

Veerkamp, W. J. J. (1996). *Statistical inference for adaptive testing*. Internal report. Enschede, The Netherlands: University of Twente.

Yen, W. M., Burket, G. R., & Sykes, R. C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model. *Psychometrika*, *56*, 39–54. doi:10.1007/BF02294584

Author

DAVID MAGIS is a research associate of the *Fonds de la Recherche Scientifique—FNRS* (Belgium), Department of Education, University of Liège, Grande Traverse 12, B-4000 Liège, Belgium, and a research fellow of the KU Leuven, Belgium; e-mail: david.magis@ulg.ac.be. His research interests include statistical applications in psychometrics and educational measurement.

Manuscript received March 26, 2014
First revision received July 23, 2014
Second revision received August 21, 2014
Accepted August 27, 2014

Downloaded from http://jebs.aera.net at KU Leuven University Library on March 7, 2016