Mesure et évaluation en éducation

Item response theory in educational assessment and evaluation

Cees A.W. Glas

Volume 31, numéro 2, 2008

URI : https://id.erudit.org/iderudit/1025005ar
DOI : https://doi.org/10.7202/1025005ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Glas, C. A. (2008). Item response theory in educational assessment and evaluation. *Mesure et évaluation en éducation*, *31*(2), 19–34. https://doi.org/10.7202/1025005ar

Résumé de l'article

La théorie de réponse à l'item (TRI) fournit un cadre utile et théoriquement bien fondé pour la mesure en éducation. Elle soutient des activités telles que la construction d'instruments de mesure, les procédures de mise en relation et de vérification d'équivalence des mesures, l'évaluation du biais d'un test et le fonctionnement différentiel d'items. Elle prévoit la base pour des banques d'items et des designs flexibles pour l'administration d'un test, comme les méthodes d'échantillonnage multicritérié, « *flexi-level testing* », et la méthode du test adaptatif par ordinateur. Tout d'abord, une brève introduction aux principes de modèles TRI est donnée. Les modèles discutés concernent des items dichotomiques (items qui sont corrects ou incorrects) et des items polytomiques (items à un crédit partiel, comme la plupart des questions ouvertes et questions de l'évaluation des compétences). Deuxièmement, on montre comment un modèle de mesure TRI peut être amélioré en utilisant un modèle structurel, par exemple, un modèle d'analyse de la variance, pour établir un lien entre les données provenant de tests pour mesurer le rendement et la capacité des élèves à des variables, tels leur statut socio-économique, leur niveau d'intelligence ou leur capital culturel, et à des variables caractérisant l'école et le système scolaire. Deux applications sont présentées. La première se rapporte aux procédures de type mise en parallèle (*equating* et *linking*), et la seconde à une combinaison d'un modèle de mesure TRI et d'un modèle linéaire multiniveaux utilisé dans la recherche relative à l'efficacité de l'école.

Item response theory in educational assessment and evaluation

Cees A.W. Glas
University of Twente

Key words: Educational assessment, educational evaluation, item response theory, one-parameter logistic model, school effectiveness research, test equating, two-parameter logistic model

Item response theory provides a useful and theoretically well-founded framework for educational measurement. It supports such activities as the construction of measurement instruments, linking and equating measurements, and evaluation of test bias and differential item functioning. It further provides underpinnings for item banking and flexible test administration designs, such as multiple matrix sampling, flexi-level testing, and computerized adaptive testing. First, a concise introduction to the principles of IRT models is given. The models discussed pertain to dichotomous items (items that are scored as either correct or incorrect) and polytomous items (items with partial credit scoring, such as most types of open-ended questions and performance assessments). Second, it is shown how an IRT measurement model can be enhanced with a structural model, such as, for instance, an analysis of variance model, to relate data from achievement and ability tests to students' background variables, such as socio-economic status, intelligence or cultural capital, to school variables, and to features of the schooling system. Two applications are presented. The first one pertains to equating and linking of assessments, and the second one to a combination of an IRT measurement model and a multilevel linear model useful in school effectiveness research.

Mots clés: Évaluation de l'éducation, théorie de réponse à l'item, modèle logistique à un paramètre, recherche sur les «écoles efficaces», test equating, modèle logistique à deux paramètres

La théorie de réponse à l'item (TRI) fournit un cadre utile et théoriquement bien fondé pour la mesure en éducation. Elle soutient des activités telles que la construction d'instruments de mesure, les procédures de mise en relation et de vérification d'équivalence des mesures, l'évaluation du biais d'un test et le fonctionnement différentiel d'items. Elle prévoit la base pour des banques d'items et des designs flexibles pour l'administration d'un test, comme les méthodes d'échantillonnage multicritérié, «flexi-level testing», et la méthode du test adaptatif par

ordinateur. Tout d'abord, une brève introduction aux principes de modèles TRI est donnée. Les modèles discutés concernent des items dichotomiques (items qui sont corrects ou incorrects) et des items polytomiques (items à un crédit partiel, comme la plupart des questions ouvertes et questions de l'évaluation des compétences). Deuxièmement, on montre comment un modèle de mesure TRI peut être amélioré en utilisant un modèle structurel, par exemple, un modèle d'analyse de la variance, pour établir un lien entre les données provenant de tests pour mesurer le rendement et la capacité des élèves à des variables, tels leur statut socio-économique, leur niveau d'intelligence ou leur capital culturel, et à des variables caractérisant l'école et le système scolaire. Deux applications sont présentées. La première se rapporte aux procédures de type mise en parallèle (equating et linking), et la seconde à une combinaison d'un modèle de mesure TRI et d'un modèle linéaire multiniveaux utilisé dans la recherche relative à l'efficacité de l'école.

Palavras-chave: Avaliação da educação, teoria da resposta ao item, modelo logístico de um parâmetro, investigação sobre as "escolas eficazes", test equating, modelo de dois parâmetros.

A teoria de resposta ao item (TRI) fornece um quadro útil e teoricamente bem fundamentado para a medida em educação. Sustenta actividades como a construção de instrumentos de medida, os procedimentos de relacionamento e de verificação de equivalência de medidas, avaliação do desvio de um teste e o funcionamento diferencial de itens. Prevê a base para os bancos de itens e desenhos flexíveis para a administração de um teste, como os métodos de amostragem multicriterial, "flexi-level testing" e o método do teste adaptativo por computador. Antes de mais, é dada uma breve introdução aos princípios dos modelos TRI. Os modelos discutidos dizem respeito aos itens dicotómicos (itens que são correctos ou incorrectos) e a itens politómicos (itens de crédito parcial, como a maior parte das perguntas abertas e das perguntas de avaliação de competências). Em segundo lugar, mostra-se como um modelo de medida pode ser melhorado utilizando um modelo estrutural, por exemplo, um modelo de análise da variância, para relacionar os dados provenientes de testes para medir o rendimento e a capacidade dos alunos com variáveis, tais como o seu estatuto socio-económico, o seu nível de inteligência ou o seu capital cultural e com variáveis que caracterizam a escola e o sistema escolar. Apresentam-se duas aplicações. A primeira está relacionada com procedimentos do tipo colocar em paralelo (equating et linking), e a segunda é uma combinação de um modelo de medida TRI com um modelo linear multinível utilizado na investigação relativa à eficácia da escola.

Author's note – All correspondence should be addressed to Prof. Dr. Cees A. W. Glas, Department of Research Methodology, Measurement, and Data Analysis, Faculty of Behavioral Science, University of Twente P.O. Box 217, 7500 AE Enschede, the Netherlands [C.A.W.Glas@gw.utwente.nl].

Introduction

Educational assessment addresses such issues as the reliability and validity of tests and examinations, linking and equating measurements, and the evaluation of test bias and differential item functioning. Educational evaluation addresses a whole range of issues from the micro-level (teaching, instrumentation, curriculum) to the macro-level (school-effectiveness research and large scale attainment studies, such as TIMSS and PISA). The statistical theory for educational assessment and evaluation seems to have two traditions: the tradition of classical test theory (CTT) and the tradition of items response theory (IRT). Though the roots of CTT go as far back as Pearson (1904, 1907), Gulliksen's standard work *Theory of mental tests* (Gulliksen, 1950) can be seen as the first comprehensive axiomatic statement of CCT. An important extension allowing for further elaboration of the sources of unreliability in test scores has become known as generalizability theory (Cardinet, 1997; Cardinet, Tourneur & Allal, 1976, 1981; Cronbach, Glaser, Nanda & Rajaratnam, 1972). Another approach closely related to CCT is multilevel modeling (Goldstein, 1986) which is, for instance, much used in school effectiveness research. The first formulations of IRT were published during and after the Second World War (Lawley, 1943, 1944; Lord, 1952, 1953) but the most influential contributions were made some time later by Rasch (1960), Birnbaum (1968), Bock (1972) and Lord (1980). Finally, somewhere positioned between CCT and IRT are the developments in latent variable modeling such as factor analyses and linear equation modeling (see, for instance, Jöreskog, 1970). These various approaches often seem unrelated. On the other hand, several authors have pointed at the connections. For instance, Takane and de Leeuw (1987) show that factor analyses and certain versions of multilevel IRT are completely equivalent. Also in the present article the connections between the various approaches will be emphasized.

To do this, we partition the statistical model for observations into two components: a measurement model and a structural model. Suppose that we collect observations y_{ik} of persons $i = 1,...,N$ to items $k = 1,...,K$. At this moment we make no assumptions about the type of responses yet, so the responses may be either discrete or continuous. In an IRT model, it is assumed that the model for a response of a person i to an item k depends both on person parameters θ_i, and on item parameters λ_k. This response has a distribution or density function $\rho(y_{ik}|\theta_i,\lambda_k)$. This part of the model is the measurement model. Added to this is a so-called structural model. Usually, the structural model is defined on the person parameters θ_i, but the model can also be

defined on the item parameters λ_k. Only the first case is treated here, examples of the second case can, for instance, be found in Fischer (1983) or Glas and van der Linden (2003). The likelihood function is given by

$$L = \prod_{i=1}^{N}\prod_{k=1}^{K} p(y_{ik} \mid \theta_i, \lambda_k)\, g(\theta_i; \beta, \Sigma, x_i) \ , \tag{1}$$

where $g(\theta_i; \beta, \Sigma, x_i)$ is the normal density function of a linear model $\theta_i = \beta x_i + \varepsilon_i$. So x_i are observed covariates, β are regression parameters and ε_i is an error term with covariance matrix Σ. Of course, θ_i might be a scalar parameter and the variance of the error term then becomes σ^2. The structural part can be an analysis of variance model, a regression model, a factor analysis model and even a structural equation model. In the present article, we present applications of this framework to educational testing. But first a number of IRT models will be discussed.

Measurement error models

In this article, the basic idea of IRT will be introduced by an example that has been around in educational measurement courses in the Netherlands for a very long time. The example was used in courses by well-known Dutch psychometricians such as Klaas Sijtsma, Henk Kelderman, Wim van der Linden and Rob Meijer, but there is uncertainty about who created the example. In any case, the example illustrates the basic idea of IRT so well, that it would be a pity not to present it. The example is measuring body height with a questionnaire. The original questionnaire consists of 30 items, 8 of which are given in Table 1 as an example.

Table 1
Items for measuring body height with a questionnaire

1	I bump my head quite often
2	For school pictures I was always asked to stand in the first row
3	In bed, I often suffer from cold feet
4	When walking down the stairs, I often take two steps at a time
5	I think I would do well in a basket ball team
6	As a police officer, I would not make much of an impression
7	In most cars I sit uncomfortably
8	I literally look up to most of my friends
9	*Etc.*

All the items have two response categories: "agree" or "disagree". All items are related to body height, but the orientation of the items is not always in the same direction. Note that a positive response to the first item is an indication of a tall stature, while a positive response to the second item is an indication of a short stature. If all the item responses are rescaled in the same direction, that is, as indications of height, then it might be reasonable to assume that the probability of endorsing an item increases with height. Further, we could try to position the items and respondents on some scale, in such a way that the scale values of respondents reflect the number of items they endorse and the scale values of items reflect the number of positive responses they attract. Consider the example of Figure 1. Suppose that items below Jim's scale value, say θ_{Jim}, are items where the probability that they are endorsed by Jim is high, say higher than 0.5. Jo dominates more items, and it is expected that she produces a higher number of positive responses. So she is taller. An analogous reasoning holds for the items. Item 6 is only dominated by Jo, so obviously you have to be quite tall to produce a positive response to that item. On the other hand, Item 3 is dominated by all three respondents so if you don't dominate that item, you must be quite short.

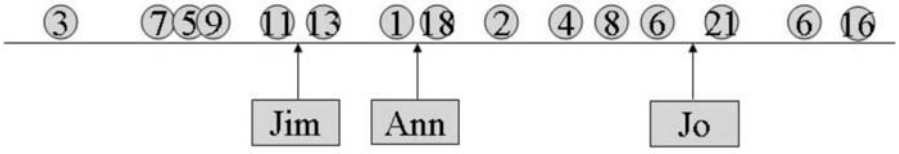


Figure 1. ***Ordering of persons and items on a latent height scale***

The questionnaire does not measure body height directly. However, from the responses and the estimated response probabilities that we estimate from the responses, we might infer that there is an unobserved dimension that produces some form of regularity in the data. This dimension is called the latent variable and in the present case, the latent variable can be identified as body height.

Estimation of the latent scale values requires a model. We define a response variable y_{ik} for a student i and an item k. In the present case, there are two possible outcomes defined by

$$y_{ik} = \begin{cases} 1 & \text{if person } i \text{ responded positive to item } k \\ 0 & \text{if this is not the case.} \end{cases}$$

A simple model where every respondent is represented by one latent parameter (body height) and every item is represented by one single item parameter will be considered first. Whether the model actually fits available data must be investigated using statistical tests of model fit. In the present case, we choose an IRT model known as the 1-parameter logistic model (1PLM) or Rasch model (Rasch, 1960). In the Rasch model, the probability of a positive response is given by

$$p(Y_i = 1 \mid \theta_i, b_k) \;=\; P_k(\theta_i) \;=\; \frac{\exp(\theta_i - b_k)}{1 + \exp(\theta_i - b_k)} \quad . \tag{2}$$

The probability of a positive response as a function of ability, $P_k(\theta)$, is the so-called item response function of item k. Two item response curves are shown in Figure 2. The x-axis is the latent continuum θ and the y-axis is the probability of a positive response.

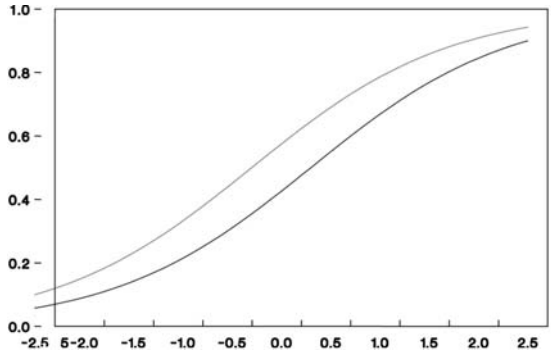


Figure 2. ***Response curves for two items in the Rasch model***

The Rasch model was developed to analyze educational testing data. Therefore, the latent variable θ is usually called ability and the item parameters b_k are usually called item difficulties. Fischer (1974) shows that the model can be derived from a number of assumptions. One is that the number-correct scores of the respondents and the numbers of correct responses given to the items are sufficient statistics for one-dimensional ability parameters θ_i and one-dimensional item parameters b_k. That is, these statistics contain all the information necessary to estimate these parameters. With the assumption of independence between responses given the model parameters and the assumption that the response functions are continuous, with the upper and lower limit going to zero and one, respectively, the 1PLM model follows.

One of the properties of the model is that the item response curves are shifted curves that don't intersect. This model may not hold for actual data. For instance, because students can guess the correct answer to a multiple choice item, the probability of a correct response does not go to zero for low ability levels. To model this more parameters are needed. In the 3PLM (Birnbaum, 1968), the probability of a correct response depends on three item parameters, a_k, b_k and c_k, which are called the discrimination, difficulty and guessing parameter, respectively. The model is given by

$$P_k(\theta_i) = c_k + (1-c_k)\frac{\exp(a_k(\theta_i - b_k))}{1+\exp(a_k(\theta_i - b_k))} \quad . \tag{3}$$

The 2PLM follows by setting the guessing parameter equal to zero. Details on estimating and testing the models can, for instance, be found in Bock and Aitkin (1981), Fischer and Molenaar (1995) and Glas and Suárez-Falćon (2003).

IRT models are generalized in many directions. For instance, models are available where the response is not dichotomous but polytomous (Masters, 1982; Muraki, 1992; Samejima, 1969) or continuous (such as response times, van der Linden, 2006). In this article, however, another generalization is discussed where the ability is not one-dimensional but multidimensional. Multidimensional IRT (MIRT) models for dichotomously scored items were first presented by McDonald (1967) who used a normal ogive to describe the probability of a correct response. Formulations based on logistic probability models were developed by Reckase (1985). MIRT models fit the framework of Formula (1) in that they consist of an IRT measurement model and a structural model. For the dichotomous case, the probability of a correct response is given by

$$P_k(\theta_i) = c_k + (1-c_k)\frac{\exp\left(\sum_{q=1}^{Q} a_{kq}\theta_{iq} - b_k\right)}{1+\exp\left(\sum_{q=1}^{Q} a_{kq}\theta_{iq} - b_k\right)} \quad , \tag{4}$$

where the parameters θ_{iq} ($q = 1,...,Q$) are the ability parameters (or factor scores) of student i, b_k is the difficulty of item k, and a_{kq} ($q = 1,...,Q$) are the factor loadings expressing the relative importance of the Q ability dimensions for giving a correct response to item k. For the structural model, it is assumed that the ability parameters θ_{iq}, have a Q-variate normal distribution with a

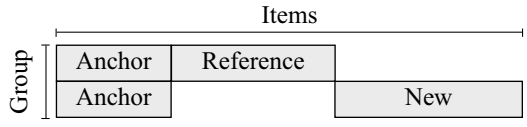
mean-vector $\boldsymbol{\mu}$ with the elements μ_i and a covariance matrix $\boldsymbol{\Sigma}$. The relative importance of the ability dimensions for the responses to specific items is modeled by item-specific loadings a_{kq} and the relation between the ability dimensions in some population of respondents is modeled by the correlation between the ability dimensions. The model can be estimated and tested by various maximum likelihood and Bayesian methods (Béguin & Glas, 2001; Bock, Gibbons & Muraki, 1988; Muthén, 1984).

Test equating and linking of assessments

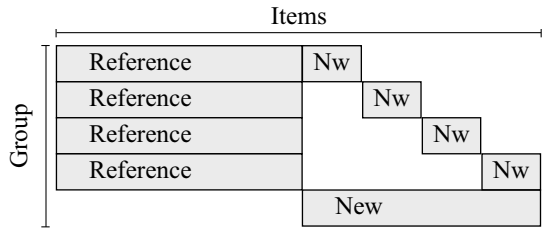
One of the important features of IRT is the possibility of analyzing so-called incomplete designs where different groups of persons have responded to different sets of items. As an application we present the equating procedure for the national examinations at the end of secondary education in the Netherlands. The grade level that students achieve on these examinations is an important component for streaming to tertiary education. Although much attention is given to producing examinations of equivalent substantive content and difficulty, research commissioned by the Inspection of Secondary Education in the Netherlands has shown that the difficulty of examinations can still fluctuate significantly over the years. This research has also shown that the proficiency level of the examinees fluctuates significantly over time. Therefore, a test equating procedure has been developed for setting the cut-off scores of examinations in such a way that differences in difficulty of examinations are taken into account. First, a latent cut-off point is set on the latent ability scale. Then the observed cut-off points on the examinations are the expected scores given the latent cut-off point. These observed cut-off points are different due to the differences in the difficulties of the examinations. The reference examination was such that its quality and difficulty presented a suitable reference point. We will discuss three equating designs that were considered for the equating procedure, as shown in Figure 3.

In the first design, every year, some weeks before the examination takes place, the students are given an anchor test covering material comparable to the content matter of the examinations. The design is depicted in the first panel of Figure 3. The figure is a graphical representation of a data matrix with persons as rows and items as columns. The shaded area denotes which items have been administered to which students. The rest is unobserved. The size of the area has no significance, that is, it does not reflect the sample sizes. Field trials showed that the problem of this design is that there are differences

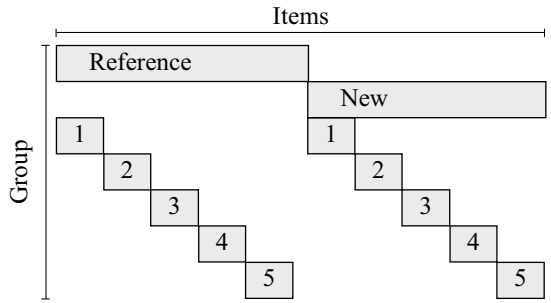
in response behavior between the administration of the anchor test and the actual examination. Firstly, the level of proficiency of the students changes during the weeks between taking the anchor test and the examination. This creates a model violation, because the person parameters in an IRT model are supposed to be constant. Further, differences in ability were accompanied by differences in item parameters which created an additional model violation. One reason was that there was a lot of guessing on the anchor test so the same set item parameters did not properly describe response behavior on the two occasions. This, of course, does not disqualify the anchor test approach in general. In many situations the gain in proficiency will be negligible and there will be no change in the item parameters. However, for the present application the decision was not to choose this approach.



Panel 1 : Anchor Item Design



Panel 2 : Pretest Design



Panel 3 : Post-test Design

Figure 3. ***Item administration designs for equating examinations***

The second design which was considered is depicted in the second panel of Figure 3. The design shown is used in the standard-setting procedure for the Swedish Scholastic Aptitude Test (SweSAT, Emons, 1998). In this design the students taking the reference examination also respond to items of a future examination. The additional items have no relevance for the final mark obtained by these students and the students are not told in advance which items do not belong to the actual examination. The strong point of the SweSAT pretest design is that the motivation of the students used in the pretest is guaranteed.

The third design depicted in Figure 3 is the design that was actually chosen. In this design the linking groups consist of students not participating in the actual examinations. They respond to items of the old and the new examination. The linking groups were presented their tests directly after the new examination was administered. Linking groups are sampled from another stream of secondary education and the design is such that the linking groups together cover all items of the two examinations. One of the concerns when planning the design is to avoid order effects. If, for instance, items from the new examination are always last, declining concentration and fatigue may result in lowering performance, so that the items of the new examination appear more difficult.

The method has been in operation for a decade now, and the results prove very satisfactory.

Multilevel IRT

In educational research, elementary units are clustered in higher-level units. A well-known example is students nested within classrooms, classrooms within schools, schools within districts and so on. Multilevel models have been developed to take the resulting hierarchical structure into account, mostly by using regression-type models with random coefficients (Goldstein, 1986). However, if variables in these multilevel models contain large measurement errors, the resulting statistical inferences can be very misleading. A solution is the so-called multilevel IRT model (MLIRT, Fox & Glas, 2001, 2003). The dependent variables are observed item scores y_{ijk}, where the index i ($i = 1,\ldots,n_j$) signifies the respondents, the index j ($j = 1,\ldots,J$) signifies the level two clusters, say the schools, and the index k ($k = 1,\ldots,K$) signifies the items. The first level of the structural multilevel model is formulated as

$$\theta_{ij} = \beta_{0j} + \beta_{1j}x_{1ij},\ldots,+\beta_{q'j}x_{q'ij} + \beta_{(q'+1)j}\xi_{(q'+1)ij} +,\ldots,+\beta_{Qj}\xi_{Qij} + e_{ij}$$

where the covariates x_{qij} ($q = 1,\ldots,q'$) are manifest predictors and the covariates ξ_{qij} ($q = q'+1,\ldots,Q$) are latent predictors. Finally, e_{ij} are independent and normally distributed error variables with mean zero and variance σ^2. In general, it is assumed that the regression coefficients β_{qj} are random over groups, but they can also be fixed parameters. In that case, $\beta_{qj} = \beta_q$ for all j. The Level 2 model for the random coefficients is given by

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}z_{1qj} +,\ldots,\gamma_{qs'}z_{s'qj} + \gamma_{q(s'+1)}\zeta_{(s'+1)qj} +,\ldots,+\gamma_{qS}\zeta_{Sqj} + u_{qj}$$

where z_{sqj} ($s = 1,\ldots,s'$) and ξ_{sqj} ($s = s'+1,\ldots,S$) are manifest and latent predictors, respectively. Further, u_{qj} are error variables which are assumed independent over j and have a Q-variate normal distribution with a mean equal to zero and a covariance matrix T.

An example of a MLIRT model is given in the path diagram in Figure 4. The structural model is presented in the square box in the middle. The structural model has two levels: the upper part of the box gives the first level (a within-schools model), and the lower part of the box gives the second level (a between-schools model). The dependent variable θ_{ij}, say math ability, is measured by three items. The responses to these items are modeled by the 2PLM with item parameters a_k and b_k, $k=1,\ldots,3$. The measurement error models are presented by the ellipses. Both levels have three independent variables: two are observed directly, and one is a latent variable with three binary observed variables. For instance, on the first level, X_{1ij} could be gender, X_{2ij} could be age, and ξ_{3ij} could be intelligence as measured by an intelligence test. On the second level, Z_{10j} could be school size, Z_{20j} could be the school budget and ζ_{30j} could be a school's pedagogical climate measured by questionnaire. It is assumed that only the intercept β_{0j} is random, so the Level 2 predictors are only related to this random intercept. The parameters in the MLIRT model are estimated in a Bayesian framework (Fox & Glas, 2001, 2003).

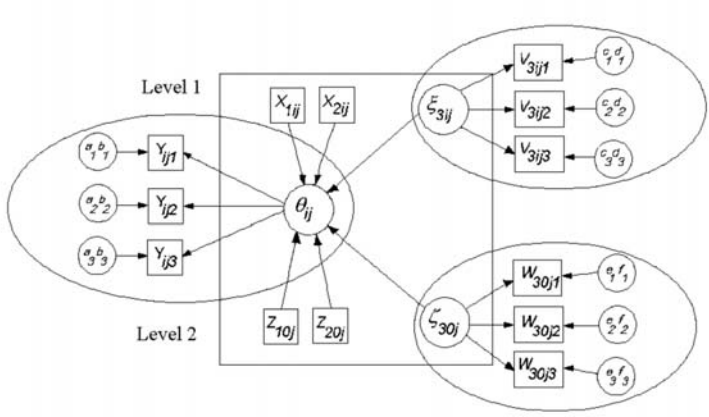


Figure 4. ***Path diagram of a multilevel IRT model***

As an example, consider an application reported Shalabi (2002). The data were a cluster sample of 3,384 grade 7 students in 119 schools. At the student level the variables were gender (0 = male, 1 = female), SES (with two indicators: the father's and mother's education, scores ranged from 0 to 8), and IQ (range from 0 to 80). At the school level: leadership (measured by a scale consisting of 25 five-point Likert items, administered to the school teachers), school climate (measured by a scale consisting of 23 five-point Likert items) and mean IQ (the IQ scores aggregated at school level). The item scores for the leadership and climate variables were recoded to dichotomous variables. The dependent variable was a mathematics achievement test consisting of 50 items. The 2PLM was used to model the responses on the leadership and school climate questionnaire and the mathematics test. For a complete description of all analyses, one is referred to Shalabi (2002); here only the estimates of the final model are given as an example. The model is given by

$$\theta_{ij} = \beta_{0j} + \beta_1 \text{SES}_{ij} + \beta_2 \text{Gender}_{ij} + \beta_3 \text{IQ}_{ij} + e_{ij}$$

and

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \text{Mean-IQ}_j + \gamma_{02} \text{Leadership}_j + \gamma_{03} \text{Climate}_j + u_{0j}$$

The results are given in Table 2. The estimates of the MLIRT model are compared with a traditional multilevel (ML) analysis where all variables were manifest. The observed mathematics, leadership and school climate scores were transformed in such a way that their scale was comparable to the scale used in the MLIRT model. Further, the parameters of the ML model were also

estimated with a Bayesian approach using the Gibbs sampler. The columns labeled C.I. give the 90% credibility intervals of the point estimates; they were derived from the posterior standard deviation. Note that the credibility regions of the regression coefficients do not contain zero, so all coefficient can be considered significant at the 90% level. It can be seen that the magnitudes of the fixed effects in the MLIRT model were larger than the analogous estimates in the ML model. This finding is in line with the other findings (Fox & Glas, 2001, 2003; Shalabi, 2002), which indicates that the MLIRT model has more power to detect effects in hierarchical data where some variables are measured with error.

Table 2
Estimates of the Effects of Leadership, Climate and Mean IQ.

	MLIRT estimates		ML estimates	
	Estimates	C.I.	Estimates	C.I.
γ_{00}	-1.096	-2.080 - -.211	0.873	-1.20 - -0.544
β_1	0.037	0.029 - 0.044	0.031	0.024 - 0.037
β_2	0.148	0.078 - 0.217	0.124	0.061 - 0.186
β_3	0.023	0.021 - 0.025	0.021	0.019 - 0.022
γ_{01}	0.017	0.009 - 0.043	0.014	0.004 - 0.023
γ_{02}	0.189	0.059 - 0.432	0.115	0.019 - 0.210
γ_{03}	-0.136	-0.383 - -0.087	-0.116	-0.236 - 0.004
Variance components				
τ_0^2	0.177	0.120 - 0.237	0.129	0.099 - 0.158
σ^2	0.189	0.164- 0.214	0.199	0.190 - 0.210

Conclusion

In this article, it is shown that the definition of a model that is a compound of a measurement model and a structural model can unify many much used models in educational evaluation and assessment that at first sight seem to have little connection. The scope of the article is limited, so we only scratch the surface of the possibilities. For instance, the dependency structure defined by Formula (1) is quite simple. IRT models with much more complicated dependence structures could be incorporated, such as the testlet model by Bradlow, Wainer and Wang (1999) and the models for ratings by Patz and

Junker (1999). Also on the side of the structural model, much was left out of consideration. The inclusion of a generalizability model is one of the most obvious candidates for further development. Finally, also just a few applications were discussed: test equating and a multilevel model for school effectiveness research. The combination of a rapidly expanding field (only to mention emergence of competency based testing) and the availability of a sophisticated test theory create many opportunities for developmental work in the future.

REFERENCES

Béguin, A.A., & Glas, C.A.W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika, 66*, 541-562.

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). 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.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM-algorithm. *Psychometrika, 46*, 443-459.

Bock, R.D., Gibbons, R.D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.

Cardinet, J. (1997). From classical test theory to generalizability theory: The contribution of ANOVA. *European Journal of Applied Psychology, 47*, 197-204.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement, 13*, 119-135.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement, 18*, 183-204.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Emons, W.H.M. (1998). *Nonequivalent groups IRT observed score equating: Its applicability and appropriateness for the Swedish Scholastic Aptitude Test*. Enschede, The Netherlands: Twente University.

Fischer, G.H. (1974). *Einführung in die theorie psychologischer tests: Introduction to the theory of psychological tests*. Bern: Huber.

Fischer, G.H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3-26.

Fischer, G.H., & Molenaar, I.W. (1995). *Rasch models. Their foundation, recent developments and applications*. New York, NJ: Springer.

Fox, J.P., & Glas, C.A.W. (2001). Bayesian Estimation of a Multilevel IRT Model using Gibbs Sampling. *Psychometrika, 66*, 271-288.

Fox, J.P., & Glas, C.A.W. (2003), Bayesian modeling of measurement error in predictor variables using Item Response Theory. *Psychometrika, 68,* 169-191.

Glas, C.A.W., & Suárez-Falćon, J.C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*, 87-106.

Glas, C.A.W., & van der Linden, W.J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27,* 256-272.

Goldstein, H. (1986). Multilevel mixed linear models analysis using iterative generalized least squares. *Biometrika, 73,* 43-56.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NJ: Wiley.

Jöreskog, K.G. (1970). A general method for analysis of covariance structures. *Biometrika, 57,* 239-251.

Lawley, D.N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 61,* 273-287.

Lawley, D.N. (1944). The factorial analysis of multiple test items. *Proceedings of the Royal Society of Edinburgh, 62-A,* 74-82.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monograph, No. 7*.

Lord, F.M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517-548.

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

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometric Monographs, No.15.*

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16,* 159- 176.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered, categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

Patz, R.J., & Junker, B.W. (1999). Applications and extensions of MCMC in IRT: Multiple Item Types, Missing Data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* 342-366.

Pearson, K. (1904). On the laws of inheritance in man. II. On the inheritance of mental and moral character in man and its comparison with the inheritance of the physical character. *Biometrika, 3*, 131-160.

Pearson, K. (1907). Mathematical contributions to the theory of evolution. XVI. On further methods of determining correlation. Draper's Company, *Research Memoirs,* Biometrics Series IV, pp. 39. London: Cambridge University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika, Monograph Supplement, No. 17*.

Shalabi, F. (2002). *Effective schooling in the West Bank*. Doctoral thesis, Twente University.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology, 3*, 271-295.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

van der Linden, W.J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*, 181-204.