Item Response Theory

W J van der Linden, CTB/McGraw-Hill, Monterey, CA, USA

Glossary

Ability parameter – Parameter in a response model that represents the person's ability, skill, or proficiency measured by the test items.
Adaptive testing – Mode of testing with real-time ability estimation in which each item is selected to be optimal at the last estimate.
Cognitive-component models – Response models with parameters for the individual cognitive operations required to solve the test item.
Dichotomous response models – Response models for items with dichotomously scored responses, such as correct–incorrect, true–false, and agree–disagree items.
Hierarchical response models – Response models with random item and/or person parameters for which a distribution is specified.
Information function – Fisher's measure for the information in an item or test score as a function of the ability parameter.
Item parameter – Parameter in a response model that represents a property of the item, such as its difficulty and discriminating power.
Multidimensional response models – Response models with a multidimensional ability parameter.
Optimal test design – Use of techniques from mathematical programming to design tests that are optimal with respect to an objective function and meet the set of constraints representing its content specifications.
Polytomous response models – Response models for items with polytomously scored responses.
Response function – Probability of a specific response on an item as a function of the ability parameter.

Item response theory (IRT) has many roots, and it would be wrong to single out anyone as the most important. But the contributions by Louis Thurstone (1925) have been decisive in many respects. These contributions built on Binet's intelligence test developed to separate students in Paris schools that needed special education from regular underachievers (Binet and Simon, 1905) as well as earlier work in psychophysics by Fechner.

Although Binet's test has been hailed as the first to be a fully standardized test in the history of educational and psychological measurement, his most significant contribution was his idea to scale the items before using them in the test. Of course, there is no natural intelligence scale, and neither were there any points of view to look at the matter in Binet's days. Binet's solution to the problem, however, was equally simple as ingenious: he chose chronological age as the scale on which he mapped his items and students. Using an empirical pretest, he defined the scale value of an item as the age group of which 75% of its members had solved it correctly. These scale values were then used to estimate the age group for which the student's achievements on the test were representative. The age of this group was the mental age of the student.

The chronological age scale used by Binet allows for direct measurement. Thurstone's innovation was to put Binet's items on an intelligence scale that cannot be measured directly. He did so because he recognized that intelligence is an example of what is now known as a latent variable; that is, an unmeasured hypothetical construct. The scale of this variable was defined by postulating a normal distribution for each age group and inferring the scale values of the items from the response data. Thurstone also showed how to check this distributional assumption. The assumption of a normal cumulative distribution function as a response function was not new but borrowed from the earlier work on psychophysics by Fechner, who used them to describe how psychological sensations vary with the strength of experimentally manipulated physical stimuli. But Thurstone's idea to separate intelligence from age and define it as a latent variable with a scale defined by such response functions was entirely new.

The idea of response functions on a latent variable was picked up again by authors like Ferguson, Lawley, and Mosier in the 1940s (and led to much confusion between the use of the normal ogive as a definition of a population distribution and a response function on a latent variable). But we had to wait until the seminal work by Lord (1952) and Rasch (1960) until the developments really began. From a more statistical point of view, later contributions by Birnbaum (1968) were important. He replaced the normal ogive by the logistic function, introduced additional item parameters to account for guessing on items (which is typical of most educational measurements), derived maximum-likelihood estimators for the model, and showed how to assemble tests from a bank of calibrated items to meet optimal statistical specifications for their application.

The next two decades showed much research on IRT models for test items with other response formats than simple dichotomous scores as well as on newer procedures for parameter estimating and model evaluation. Especially, the development of Bayesian procedures for parameter estimation and model validation was prominent. When computers became more powerful and cheaper in the 1980s, the routines for maintaining common scales and test assembly used in most large-scale educational testing programs became IRT based. The first programs to exploit IRT to score test takers in real time and deliver computerized adaptive tests were launched in the 1990s. Nowadays, IRT models and procedures are no longer the main instruments only in the educational testing industry but are becoming increasingly popular in psychological testing as well. In addition, they have hit areas such as medical diagnosis and marketing research.

Review of Response Models

Unidimensional Logistic Models for Dichotomous Items

Due to Thurstone's pioneering work, the assumption of a normal-ogive function as response function for test items remained popular for a long time. Lord's (1952) treatment of IRT for dichotomously scored items was also based on this assumption. Let U_i denote the response to item i, where $U_i = 1$ if the response on item i is correct and $U_i = 0$ if it is incorrect. The normal-ogive model describes the probability of a correct response as a function of the test taker's ability θ:

$$p_i(\theta) = \Pr\{U_i = 1|\theta\} = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} \exp^{-z^2/2} dz. \quad [1]$$

where $\theta \in (-\infty, \infty)$ and $b_i \in (-\infty, \infty)$ and $a_i \in (0, \infty)$ are parameters for the difficulty and discriminating power of item i.

Observe that the model has its parameter structure in the upper limit of the integral. For this reason, it was rather intractable at the time, and hardly used in routine applications in testing. The model also seemed to ignore the possibility of guessing on test items. These two points were addressed in Birnbaum's (1968) model, which is now generally known as the three-parameter logistic (3PL):

$$p_i(\theta) = \Pr\{U_i = 1|\theta\} = c_i + (1 - c_i)\frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))} \quad [2]$$

where additional parameter $c_i \in [0, 1]$ represents the height of a lower asymptote for the probability of a correct response. The asymptote is approached for $\theta \to -\infty$. This limit is assumed to represent random guessing without any knowledge.

The two models have no fixed unit and origin for the scale of θ. In practice, we therefore fix the scale following a practical convention. All later models in this article need comparable constraints to become identifiable. For an appropriate choice of scale unit, the models in eqns [1] and [2] predict success probabilities for identical sets of parameter values that are virtually indistinguishable from each other.

A graphical example of the logistic response function for a set of item parameter values is given in **Figure 1**. Difficulty parameter $b_i = 1$ controls the location of response function along the θ scale. A more difficult item has its location to the right of $\theta = 1$ and requires a higher ability to give the same probability of success. Discrimination parameter $a_i = 1.4$ controls the slope of the response function. A more discriminating item has a response function with a steeper slope than in **Figure 1**. Finally, a value of 0.23 for discrimination parameter c_i indicates the height of the lower asymptote for the response function. Typically, for multiple-choice items, the estimated values for this guessing parameter are close to the reciprocal of their number of response alternatives.

The two-parameter logistic (2PL) model is obtained if we put $c_i = 0$ in eqn [2]. The result is the logistic analog of eqn [1]. If we also assume equal values for the discrimination parameter (that is, $a_i = a$ for all i), the Rasch (1960) model is obtained. This model belongs to the exponential family of distributions in statistics and is statistically quite tractable.

Due to its parameter structure, the 3PL model is flexible and has been shown to fit large pools of items written for the same content domain in educational and psychological testing. In fact, it has become the standard of the testing industry for tests with dichotomous items

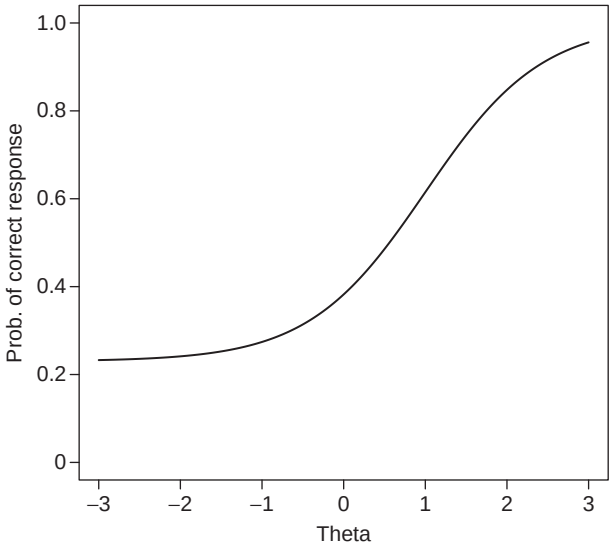

Figure 1 Example of a response function for the 3-parameter logistic model in eqn [2].

that measure a unidimensional ability. For more narrowly defined domains of educational items, or psychological tests of well-defined constructs, the Rasch model becomes an attractive alternative.

Models for Polytomous Items

Test items have a polytomous format when the responses are scored in more than two categories. Such models have a reponse function for each different category. **Figure 2** shows a typical set of response functions for an item with five different categories for one of the polytomous response models below.

Although the response categories of dichotomous items typically have a fixed order (e.g., correct–incorrect; true–false), this does not necessarily hold for polytomous items. Polytomous items have a nominal response format if their response categories can be classified but an *a priori* order between them does not exist. The nominal response model below is appropriate for items with this format. If an *a priori* ordering of the categories does exist, a graded-response model or partial-credit model should be chosen. The two models differ in the problem-solving process that is assumed to lead to the responses.

Models for a graded-response format are more generally known as cumulative models in ordinal categorical data analysis. The partial-credit models below are known as adjacent-category models. These two options do not exhaust all possibilities, a more comprehensive review of different polytomous response formats and IRT models is given in Mellenbergh (1994).

Nominal-response model

Response variable U_i is now assumed to have possible values $h = 1, \ldots, m_i > 2$. Observe that different items in the same test may have different numbers of possible responses. According to the nominal-response model (Bock, 1972), the response function for category h is

$$p_{ih}(\theta) = \Pr\{U_i = h|\theta\} \quad [3]$$

$$= \frac{\exp(a_{ih}(\theta - b_{ih}))}{\sum_{h=1}^{m_i} \exp(a_{ih}(\theta - b_{ih}))}. \quad [4]$$

Parameters a_{ih} and b_{ih} maintain their interpretation as a discrimination and difficulty parameter. **Figure 2** illustrates that b_{ih} represents the location of the response function for category h along the ability scale, where the location is defined as the value of θ at which the function shows its largest change in curvature. The value of the discrimination parameters a_{ih} is proportional to this change.

In spite of the nominal response format, the values for the parameters b_{ih}, $h = 1, \ldots, m_i$, do imply an order for the response categories. However, the actual order is only known when these parameters are estimated; it is thus empirical, not *a priori*. This feature makes the model less suitable for educational measurement, where performances on test items can always be ordered from worse to better in advance. Another reason why the model in eqn [4] may be less appropriate for this application is that it does not allow explicitly for guessing. A version of the nominal response model that does allow for guessing is given in Thissen and Steinberg (1997).

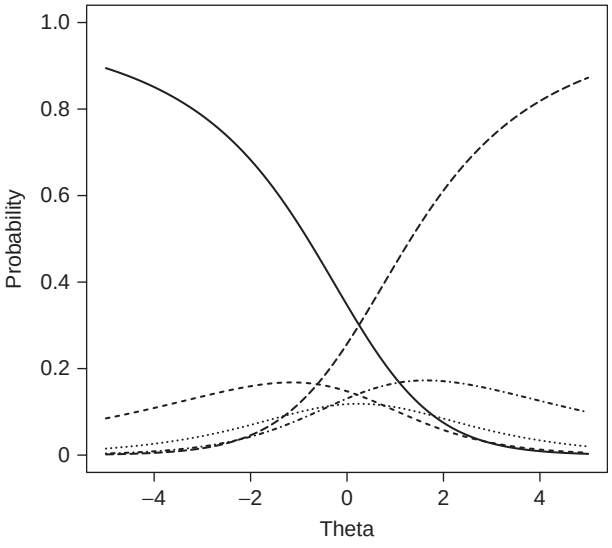

Figure 2 Example of response functions for a polytomous response model with five categories.

Graded-response model

Suppose index h reflects an *a priori* ordering of the response categories. The graded-response model (Samejima, 1969) addresses the probabilities of the compound events $U_i \geq h$,

$$P_{ih}(\theta) = \begin{cases} 1 & \text{for} h = 1 \\ \Pr\{U_i \geq h|\theta\} & \text{for} h = 2, \ldots, m_i \\ 0 & \text{for} h > m_i \end{cases} \quad [5]$$

as a function of ability parameter θ. The more interesting probabilities are those for $h = 2, \ldots, m_i$. They increase monotonically with θ because the probability of responding in any of these categories or higher goes to one if $\theta \to \infty$.

A typical choice for $P_{ih}(\theta)$ for $h = 2, \ldots, m$ in eqn [5] is from the logistic functions in eqn [2] for $c_i = 0$. If the parameters a_i are free, the result is known as the nonhomogeneous case of the model. If we set $a_i = 1$ for all i, a more stringent version of the graded-response model is obtained, which is known as its homogenous case.

The response functions for the individual categories $h = 2, \ldots, m_i$ can be derived from eqn [5] as

$$p_{ih}(\theta) = P_{ih}(\theta) - P_{i(h+1)}(\theta). \quad [6]$$

The shape of these response functions may not differ much from those for the nominal-response model.

The most distinctive feature of these functions, however, is that they reflect an *a priori* order, which is specified by the test specialist when assigning the labels $h = 1, \ldots, m_i$ to the categories. These labels determine the way in which the differences in eqn [6] are calculated.

Partial-credit models

These models derive their name from an increasing credit given to the responses $h = 1, \ldots, m_i$. All models are based on a sequential response process in which the test taker is assumed to compare two adjacent categories at a time and decide to prefer one category over the other with a certain probability.

The first version of the partial-credit model (Masters, 1982) was based on the assumption that the probability of preferring response h relative to $h-1$ can be represented by the Rasch model. Upon some probability calculus, this assumption can be shown to lead to the following response functions for $h = 1, \ldots, m_i$:

$$p_{ih}(\theta) = \frac{\exp\left(\sum_{k=1}^{h}(\theta - b_{ik})\right)}{\sum_{h=1}^{m_i}\exp\left(\sum_{k=1}^{h}(\theta - b_{ik})\right)}. \qquad [7]$$

A generalized version of the model exists in which the probabilities of adjacent responses are based on the 2PL model with free discrimination parameters a_i (Muraki, 1992). The model therefore has parameter structures $(\theta - b_{ih})$ extended to $a_i(\theta - b_{ih})$.

Further, if we adopt a common number of categories m for all items, the generalized partial-credit model can be specialized to a model for a set of rating scales of the Likert type. Likert scales are well known in attitude measurement, where they are used to ask subjects to evaluate a set of attitude statements using scales with common categories such as strongly agree, agree, neutral, disagree, and strongly disagree. The steps necessary to obtain the rating scale model steps are: (1) decomposing the parameters b_{ih} additively as $b_i + d_k$, with b_i a location parameter for the entire item and d_k a threshold parameter for kth category on the Likert scale, and (2) constraining the discrimination parameters to special known constants. This rating scale model was introduced by Andersen (1977) and Andrich (1978).

Multidimensional Models

Test items may measure more than one ability, for example, a verbal ability in addition to the ability in another domain of achievement that is tested more explicitly. If this happens, the previous models have to be extended by more ability parameter. For the 3PL model in eqn [2], the extension with a second ability parameter may lead to response functions

$$p_i(\theta_1, \theta_2) = c_i + (1 - c_i)\frac{e^{a_{i1}\theta_1 + a_{i2}\theta_2 - b_i}}{1 + e^{a_{i1}\theta_1 + a_{i2}\theta_2 - b_i}}. \qquad [8]$$

This model defines the probability of a correct response as a function of (θ_1, θ_2). It has two discrimination parameters, a_{i1} and a_{i2}, which control the slope of the response surface along θ_1 and θ_2, respectively. But it has only one (generalized) difficulty parameter b_i; a similar model with two parameters $b_{i1} - b_{i2}$ would be nonidentifiable version of it. If $\theta_1, \theta_2 \to -\infty$, the probability of a correct response goes to c_i. The role of these parameters becomes clear if we view the graphical example of the response surface in **Figure 3** as a two-dimensional generalization of the response function in **Figure 1**.

Cognitive-Component Models

IRT models in this class specify the probability of success on a test item as a function of the components of the cognitive task involved in solving the item. Two main types of models have been used. Both types agree in that they decompose the difficulty parameter as a set of more basic parameters for the relevant components. They differ, however, in how to impose a structure on these basic parameters.

Let $U_{ik} = 0, 1$ denote whether or not component $k = 1, \ldots, K$ of the task involved in solving item i has been executed successfully. Suppose a correct response to item i requires a successful execution of each of these components. The probability of a correct response for this conjunctive structure is then given by the product of the probabilities for each of the components,

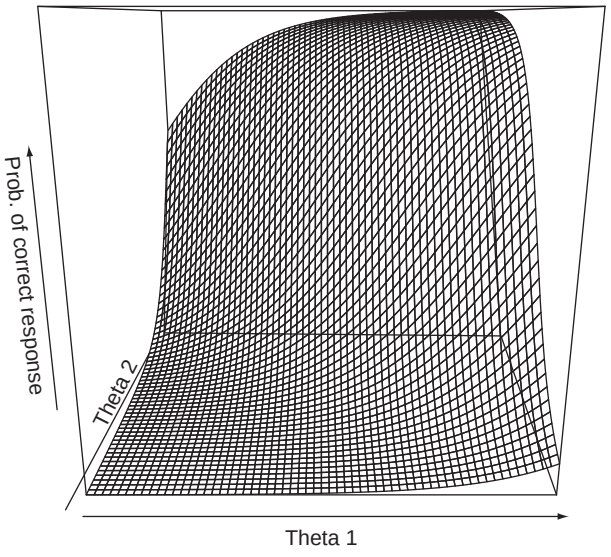

Figure 3 Example of a two-dimensional response surface for the logistic model in eqn [8].

$$\prod_{k=1}^{K} \Pr\{U_{ik}\}. \quad [9]$$

Models of this type were introduced by Embretson (1997). Her choice of success probabilities for the component tasks was the Rasch model with free ability parameters, θ_k, and difficulty parameters, β_{ik}, for $k = 1, \ldots, K$. She also introduced a guessing parameter for the case a test taker misses a component and has to guess to the answer on the item.

A different approach is offered in the linear logistic model introduced by Fischer (1983). In this model, the structure is not imposed on the probabilities of successful completion of the component tasks, such as for the conjunctive structure in eqn [9], but directly on their difficulty parameters. Assuming the same set of component tasks $k = 1, \ldots, K$ to hold for item i, the item-difficulty parameter is taken to be the following linear function of the component parameters:

$$b_i = \sum_{k=1}^{K} w_{ik}\beta_{ik} + c$$

where w_{ik} is the weight set by a content specialist to represent the impact of component difficulty β_{ik} on the difficulty of item i.

The idea of modeling underlying processes rather than the responses themselves has been a fruitful area of research. Its spirit is well captured by the name explanatory item response modeling given to the area by De Boeck and Wilson (2005).

Hierarchical Models

Hierarchical or multilevel modeling of responses is appropriate when the test takers can be assumed to be sampled from a population and the interest is in some of the features of its ability distribution. This application arises, for example, in large-scale educational assessments, where the development in the educational achievements of certain populations should be compared or followed over time. Alternatively, the interest may exist in test items that can be considered as randomly sampled, for example, from families of items generated from a set of parent items, as in the use of techniques of item cloning (Glas and van der Linden, 2003). In this case, to allow for the random sampling of items from different families, the distributions of their parameters in these families have to be modeled. Of course, it is possible to have applications in which the distributions of both the person and item parameters have to be specified.

As a first-level model, typically one of the earlier models is chosen, for example, the 3PL model for a correct response in eqn [2]. A second-level model for the ability distribution is the normal

$$\theta \sim N(\mu, \sigma^2) \quad [10]$$

with mean μ and variance σ^2.

If the items are sampled from families $f = 1, \ldots, F$, the second-level models for their distributions can be taken to be the multivariate normals

$$(a_{i_f}, b_{i_f}, c_{i_f}) \sim MVN(\mu_f, \Sigma_f) \quad [11]$$

where μ_f and Σ_f are the mean and covariance matrix for the item parameters of family f.

These models become more powerful if we are able to explain the second-level distributions of the abilities or item parameters by background variables or covariates, for example, group memberships of the test takers or structural features of the item families. This can be done by introducing linear regression structure for the means in eqn [10] or eqn [11]. Following the spirit of hierarchical linear modeling, the regression parameters can also be defined to be random at a higher level of modeling (Fox and Glas, 2001). In large-scale educational assessments, the introduction of such regression structures helps us to pinpoint differences in achievement between specific populations.

Other Models

The previous response models belong to the most important categories in the literature as well as the practice of educational measurement. But they do certainly not exhaustive the possibilities. To date, some 30–40 different response models have been proposed that are statistically tractable. The collection includes models for nonmonotone items. An item is called monotone if the function for its correct response is monotonically increasing with the ability measured by the items. This feature is typical of achievement test items but cannot be expected to hold for items in attitude scales or preference measurement, where respondents may endorse the same alternative for opposite reasons. The current review, moreover, does not include nonparametric approaches to IRT modeling or models for response times on test items. Nonparametric approaches avoid the assumption of a parametric family of response functions but try to use order assumptions with respect to probabilities of success, persons, and/or items only, which may vary in their degree of stringency. In addition, polytomous versions of nonparametric models exist (for a review, see Sijtsma and Molenaar, 2002). Response-time models have become important because of the increasing popularity of computerized delivery of educational tests with its automatic recording of the response times on the individual items. In order to use these times as an additional source of information on the persons or items, a response model with a typical IRT parameterization is necessary (van der Linden, 2006, 2009).

For more extensive introductions to larger collections of response models, the reader should consult Fischer and Molenaar (1995) and van der Linden and Hambleton (1997).

Applications in Educational Measurement

Item Calibration and Measurement

The use of IRT models in educational measurement typically consists of the stages of item calibration and measurement. During item calibration, response data are collected for a representative sample of test takers whereupon the item parameters are estimated from the data and the validity of the model is evaluated. The process of model evaluation may involve the checking of the model against the data for features such as the match between the dimensionality of the data and the ability parameters in the model, the assumption of local independence required to estimated the parameters, possible systematic differences in item parameters between relevant subpopulations (differential item functioning), and the general shape of the response functions. Additionally, it should be checked if the test takers tend to behave according to the response model. All checks take the form of statistical hypothesis testing, where the alternative hypothesis is the specific model violation against which the model is tested. For a more comprehensive treatment of these tests, see Glas and Meijer (2003) and Glas and Suarez Falcon (2003).

If the model fits and the item parameters are estimated with enough precision, the model can be used as measurement model. The test is then administered as part of operational testing and the person parameter is estimated for each test taker equating the item parameters to their estimates. The estimate is reported as the test taker's ability score. (Following a Bayesian approach, it would be more appropriate not to treat the item parameters as completely known but to account for their estimation error when estimating the person parameter. But this practice is rather unusual in educational measurement.)

Person parameters are typically estimated using the method of maximum likelihood (ML) or a Bayesian method. For the dichotomous model in eqn [2], these estimates are obtained as follows. Let $(u_1, \ldots, u_n)$ be the vector with the responses for a person on the items $i = 1, \ldots, n$ in the test that is used to measure the person's ability, θ. In ML estimation, the estimate of θ is calculated from the probability of the response vector under the model taken as a function of θ. More specifically, this function is known as the likelihood function of θ associated with the responses,

$$L(\theta|u_1, \ldots, u_n) = \prod_{i=1}^{n} p_i(\theta)^{u_i}[1 - p_i(\theta)]^{1-u_i}, \qquad [12]$$

and the ML estimate is the value of θ for which this likelihood function reaches its maximum.

In a Bayesian approach, θ is estimated by its posterior distribution; that is, its probability distribution given the responses. The distribution is obtained by assuming a prior distribution of θ and combining this with its likelihood as:

$$f(\theta|u_1, \ldots, u_n) = cL(\theta|u_1, \ldots, u_n)f(\theta), \qquad [13]$$

where $f(\theta)$ is the density of the prior distribution of θ, $f(\theta \mid u_1, \ldots, u_n)$ is the density of its posterior distribution, and c is a normalizing constant. If a point estimate is needed, the mean of the posterior distribution can be taken (expected *a posteriori* or EAP estimation).

The power of IRT for educational measurement lies in the presence of the item parameters in the expression in eqns [12] and [13] from which the ability estimate is calculated. Due to this, the estimation automatically accounts for the properties of the items that were selected in the test. For example, if the items would have been more difficult, the values of the item difficulty parameters would have been greater, and the ability estimate for the response vector would automatically have been increased to account for this. This feature has been called item-free measurement – an expression that, when taken naively, seems to capture this feature nicely but is somewhat misleading because the statistical features of the estimate (e.g., its accuracy) still depend on the chosen items.

Another way of touting the power of IRT is by pointing at the fact that it is able to produce valid item calibration or educational measurement from response data collected using research designs with missing data. The applications in the next sections capitalize on this feature. Of course, the validity of parameter estimates from incomplete designs is only guaranteed if the fact whether or not the responses are missing does not contain any direct information on their correctness. It would be wrong to leave out a portion of the responses, for instance, because they were incorrect.

Specific Applications with Missing Data

Most of the large-scale educational testing programs now use item banking. In item banking, new items are written, pretested, and calibrated continuously. If an item passes all quality checks, it is added to the item bank. At the same time, new tests are assembled from the items in the bank. This practice differs dramatically form traditional test construction, which goes through a complete cycle of item writing, pretesting, and test construction for one test at a time. Obvious advantages of item banking are more constant use of the resources, permanent access to a large stock of pretested, high-quality items for test assembly, and less vulnerability to premature leaking of a test form. Its main advantage is, however, stable score scales defined by larger pools of items. Such scales permit testing programs to

produce scores that are comparable over time even when earlier stocks of items in the bank have been retired and replenished.

Item banking is possible because when using IRT, it is no longer necessary for two persons to answer identical collections of test items to have comparable scores. Likewise, in order to compare the parameter estimates of different items, it is no longer necessary for them to be taken by the same sample of test takers. It is therefore possible to build up banks with much less restricted versions of data-collection designs than in traditional pretesting of test items. Likewise, once the items have been calibrated, we can deliberately select a subset of items for a specific measurement goal.

The idea of optimal test assembly capitalizes on the latter opportunity. One of the first to point at it was Birnbaum (1968). His idea was to translate the measurement goal into a target function for the accuracy of the test along the ability scale. The test was then assembled to match the target as closely as possible. For example, the test could be required to have a uniform target over a certain ability range when it is used for diagnostic purposes with respect to students in it, or a peaked function at a cutoff score when the goal is accurate admission decisions.

For this approach to be practical, it is necessary to have a measure of the accuracy of the test as a function of the ability θ measured by the items. A useful measure is Fisher's information in the item responses in the test, which for the 3PL model in eqn [2] can be shown to be equal to

$$I_i(\theta) = \frac{a_i^2[1 - p_i(\theta)]}{p_i(\theta)}\left[\frac{p_i(\theta) - c_i}{1 - c_i}\right]^2. \qquad [14]$$

for item i. Taken as a function of θ, the measure is generally referred to as the information function of item i. The use of this measure for optimal test assembly is motivated by two different facts: First, information functions are additive. That is, if the items $i = 1, \ldots, n$ are selected for a test, test information function $I(\theta)$ can be shown to be equal to the sum of the item information functions,

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta). \qquad [15]$$

It is thus straightforward to evaluate the effects of adding or removing an item to a test. Second, the sampling variance of the ML estimator of θ for a given test is asymptotically equal to the inverse of the information function in eqn [15]. This feature gives the use of the information function its statistical foundation. **Figure 4** illustrates the additivity of the information functions for a test of six items.

Although optimal test assembly can be illustrated graphically rather easily for a short test, the actual assembly of a test in a real-world application is no simple affair.

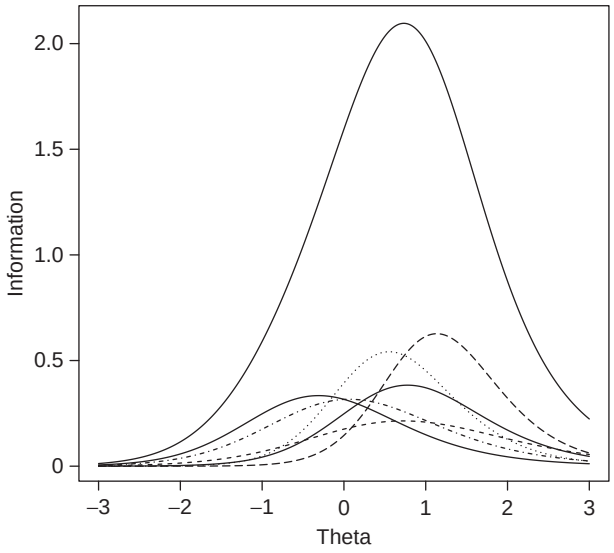

Figure 4 Example of six item information functions and their test information function.

In many applications no single tests but sets of parallel tests are to be selected. More importantly, tests always have to be assembled to sets of constraints that control their content specifications. In fact, it is not unlikely for real-world tests to be assembled to several hundreds of such constraints. Fortunately, such problems can be modeled to be a linear integer programming problem, which can easily be solved using standard commercial software. For solutions to a large variation of optimal test assembly problems with various objective functions and types of constraint, see van der Linden (2005).

In computerized adaptive testing (CAT), the items in the test are not selected simultaneously for a group of test takers but sequentially from the item pool for an individual test taker. The responses on the items are recorded and used to update the person's ability estimate in real time. Each next item in the test is selected to be optimal at the last estimate. Due to this adaptation, the ability estimate converges much faster to the person's true ability than for a traditional fixed test, even when it has been assembled to be optimal for a group of test takers.

In order to implement adaptive testing, a criterion for item selection is required. An obvious criterion is to use the information function $I_i(\theta)$ in eqn [14] and select the next item to have a maximum value for it at the last ability estimate among the items in the pool. This popular criterion of item selection is known as the maximum-information criterion. Alternatively, Bayesian item-selection criteria can be used, for example, a criterion that minimizes the variance of the posterior distribution of the test taker's ability in eqn [13].

Other practical issues that have to be addressed when implementing adaptive testing are how to impose a fixed set of content constraints on the test for different test

takers in real-time item selection, how to prevent different time pressure between test takers that get different selections of items, and how to maintain the integrity of the item pool against test takers that cheat and try to memorize and share test items. IRT-based solutions to these problems are reviewed in van der Linden (2005) and van der Linden and Glas (2010).

Bibliography

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika* **42**, 69–81.

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

Binet, A. and Simon, T. A. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectual des anormoux. *L'Anneé Psychologie* **11**, 191–336.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. and Novick, M. R. (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.

De Boeck, P. and Wilson, M. R. (2005). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer.

Embretson, S. E. (1997). Multicomponent response models. In van der Linden, W. J. and Hambleton, R. K. (eds.) *Handbook of Modern Item Response Theory*, pp 305–321. New York: Springer.

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

Fischer, G. H. and Molenaar, I. W. (eds.) (1995). *Rasch Models: Foundations, Recent Developments and Applications.* New York: Springer.

Fox, G. J. A. and Glas, C. A. W. (2001). Bayesian estimaton of a multilevel IRT model using Gibbs sampling. *Psychometrika* **66**, 271–288.

Glas, C. A. W. and Meijer, R. R. (2003). A Bayesian approach to person-fit analysis in item response theory models. *Applied Psychological Measurement* **27**, 217–233.

Glas, C. A. W. and Suarez Falcon, 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. and van der Linden, W. J. (2003). Computerized adaptive testing with item clones. *Applied Psychological Measurement* **27**, 247–261.

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

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

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin* **115**, 300–307.

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

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danisch Institute for Educational Research.

Samejima, F. (1969). Estimation of ability using a pattern of graded responses. *Psychometrika Monograph* **17**, 1–100.

Sijtsma, K. and Molenaar, I. W. (2002). *Introduction to Nonparametric Item Response Theory*. Thousand Oaks, CA: Sage.

Thissen, D. and Steinberg, L. (1997). A response model for multiple-choice items. In van der Linden, W. J. and Hambleton, R. K. (eds.) *Handbook of Modern Item Response Theory*, pp 51–65. New York: Springer.

van der Linden, W. J. (2005). *Linear Models for Optimal Test Design*. New York: Springer.

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

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational and Measurement* **46**, 247–272.

van der Linden, W. J. and Glas, C. A. W. (eds.) (2010). *Elements of Adaptive Testing.* New York: Springer.

van der Linden, W. J. and Hambleton, R. K. (1997). *Handbook of Modern Test Theory*. New York: Springer.

Further Reading

Bock, R. D. (1997). A brief history of items response theory. *Educational Measurement: Issues and Practice* **16**, 21–33.

Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.