Generalized Linear Item Response Theory

Gideon J. Mellenbergh

In this article generalized linear item response theory is discussed, which is based on the following assumptions: (a) A distribution of the responses occurs according to a given item format; (b) the item responses are explained by one continuous or nominal latent variable and p latent as well as observed variables that are continuous or nominal; (c) the responses to the different items of a test are independently distributed given the values of the explanatory variables; and (d) a monotone differentiable function g of the expected item response τ is needed such that a linear combination of the explanatory variables is a predictor of $g(\tau)$. It is shown that most of the well-known psychometric models are special cases of the generalized theory and that concepts such as differential item functioning, specific objectivity, reliability, and information can be subsumed under the generalized theory.

Linear models, which were developed in statistics, are frequently used in psychology and education for analyzing empirical data. For example, a weighted sum, the linear combination of independent variables that can be used to predict the expected value of a dependent variable, is used in selection in instances in which the weighted sum of a subject's test scores is used to predict the subject's expected criterion performance. The theory of linear models has been generalized in statistics, and it has been shown that many of the existing statistical models are special cases of the generalized theory.

Linear models are also applied in psychometrics. In the psychometric literature, the emphasis is often on the differences between the various types of models. Usually, classical test theory is contrasted with modern test theory and, in the framework of modern test theory, the one-parameter Rasch model is contrasted with the two-parameter Birnbaum model.

There are, however, a number of authors who have investigated the relationships between the different models. Masters and Wright (1984) showed that five members of the Rasch family share a common form. Moosbrugger and Müller (1982) showed that a Rasch-type model can be formulated as a special case of classical test theory. Thissen, Steinberg, Pyszczynski, and Greenberg (1983) used the one-factor model for Likert-type items and emphasized its correspondence with the Birnbaum model for dichotomous items for item and test analysis. Thissen and Steinberg (1986) presented a taxonomy of different models of modern test theory. McDonald (1982) placed the models for dichotomous items in the framework of nonlinear factor analysis. Goldstein and Wood (1989) showed that a linear model framework combined with different item response transformations can unify many psychometric models. As an extension of the work of these authors, a generalized psychometric theory is proposed that is analogous to the generalized statistical theory and also subsumes many of the existing psychometric models and concepts.

In the classical linear models used in statistics, the dependent variable (e.g., a subject's criterion performance) is predicted by one or more independent variables. The dependent variable must be measured on a continuous interval scale. The independent (predictor) variables are observed variables that can be measured on different scales. An example of a linear model that uses continuous interval predictor variables is the classical regression model. The predictor variables (e.g., test scores) must be measured on continuous scales. An example of a linear model that uses a nominal predictor variable is the classical analysis of variance model. The predictor variable is a nominal variable (e.g., the variable consisting of experimental and control groups).

In many practical situations, the dependent variable is not measured on a continuous interval scale. An example is a performance criterion that consists of success or failure in a job or a pass or fail grade on an examination, both of which are dichotomous criterion performances. Suppose test scores and classical regression are used to predict a subject's expected criterion performance. A subject's expected value on a dichotomous criterion is the probability of the subject being successful in the job or passing the examination. The expected criterion performance is a probability, which means that its lowest possible value is 0 and its highest possible value is 1. But if test scores and the regression model are used, the predicted expected criterion performance may be smaller than 0 or larger than 1. The problem is solved by using a transformation of the probability of having success such that the predicted expected criterion value can vary from minus infinity to plus infinity instead of

I thank Henk Kelderman, Wulfert P. Van den Brink, Frans J. Oort, and two reviewers for their valuable comments.

Correspondence concerning this article should be addressed to Gideon J. Mellenbergh, Faculteit der Psychologie, Universiteit van Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands.

Psychological Bulletin, 1994, Vol. 115, No. 2, 300–307
Copyright 1994 by the American Psychological Association, Inc. 0033-2909/94/$3.00

from 0 to 1. One suitable transformation is the logit transformation. The logit function transforms the probability, which can vary from 0 to 1, to a value that can vary from minus infinity to plus infinity. The resulting model is called the logistic regression model (Agresti, 1990, Section 4.2).

In statistics, the theory of linear models was extended to the theory of generalized linear models for handling these types of situations (e.g., see Dobson, 1983; McCullagh & Nelder, 1989). In the generalized theory, the expected value of the dependent variable is transformed. The transformation function depends on the measurement scale and the distribution of the dependent variable (e.g., the logit function is a suitable transformation for a binomially distributed dichotomous dependent variable). The transformed expected value is predicted by a linear combination of observed variables (e.g., the subject's logit-transformed probability of success is predicted by a weighted sum of his or her test scores).

In psychometrics, linear combinations of variables are also used to predict the expected value of an observed variable. In classical test theory, a subject's observed score is a linear combination of a true score and an error score, which means that the expected value of the subject's observed score is equal to the true score. In modern test theory, models were developed for a subject's responses to dichotomously scored test items. The subject's expected item score is the probability of that subject providing a correct answer to the item. It is assumed that a subject's response behavior is determined by one or more latent traits. However, the probability of a correct answer cannot be predicted by a linear function of one or more latent traits because the probability is in the interval from 0 to 1, whereas the predicted probability can be outside this interval. As in logistic regression, the problem can be solved by using the logit function. The logit function is used for transforming the subject's probability of providing a correct answer; the transformed probability is predicted by a linear function of one or more latent traits (e.g., see, Hambleton & Swaminathan, 1985, Section 4.2). These modern test theory models are called logistic item response models.

The classical test theory model has the same structure as the classical linear model used in statistics. The logistic item response models have the same structure as the logistic regression model used in statistics. In both cases, the (transformed) expected value of a dependent variable is predicted or explained by a linear combination of other variables. One difference between statistical and psychometric models is that in statistics the predictors are usually observed variables, whereas in psychometrics latent predictor variables are primarily used. This is also the main difference between the statistical theory of generalized linear models and the theory proposed here: generalized linear item response theory (GLIRT).

The generalized theory for item responses has several advantages. First, it explicitly shows the assumptions of and relations between different psychometric models. Therefore, a better choice of model can be made for practical applications in test construction and item analysis. Second, in different areas of psychometrics, different concepts have been developed (e.g., the concept of reliability stems from classical test theory, whereas information stems from modern test theory), most of which can be subsumed under the generalized theory (e.g., reliability and information can both be subsumed under GLIRT). Under the generalized theory, concepts that are exclusively used in one field can be used in all other fields of psychometrics. Third, not only are different concepts used in different areas of psychometrics, but also different types of models. For example, latent class models were developed and used for dichotomously scored items. These models can be formulated as special cases of the generalized theory, which means that latent class models have a broader scope than do dichotomous items. Finally, a variety of algorithms and computer programs are used in different areas of psychometrics. The generalized theory can stimulate the development of more general types of algorithms and programs.

Item Response Theory

Item response theory (IRT) consists of a series of models for describing and explaining subjects' response behavior to educational and psychological test items. The emphasis is on dichotomously scored items in achievement and performance tests (for a review, see Hambleton & Swaminathan, 1985).

The item response models for dichotomous items have a number of joint assumptions. First, a distribution of the item responses is assumed. The subject provides either a correct (agree) answer scored as 1, or an incorrect (disagree) answer scored as 0. The probability of the subject providing the correct answer and the complementary probability of the subject providing the incorrect answer constitute a distribution called the point binomial distribution. Second, it is assumed that the item responses are explained by one or more latent traits. The traits are latent variables that are measured on a continuous scale. In unidimensional latent trait models, the item responses are explained by one latent trait; in multidimensional models, they are explained by more than one trait. Third, it is assumed that, given the value(s) of the latent trait(s), the item responses are distributed independently. This is the assumption of local independence. If a test is administered to a subject, the result is a pattern of item responses (i.e., either a correct or an incorrect answer for each item of the test). The assumption of local independence states that the probability of the response pattern is equal to the product of the probabilities of the separate item responses. Fourth, a function for the regression of the item responses on the latent trait(s) is assumed.

The different item response models share the first three assumptions but differ in the fourth assumption of the item characteristic curve. The curve is the function that relates the expected item response (i.e., the probability of the subject providing a correct response) to the latent trait(s). The first item response model was developed by Guttman. In his model, the item characteristic curve is a step function: Below a certain point of the latent trait, the probability of a correct answer to the item is 0; above that point, the probability is 1. The Guttman model is a deterministic model because it only permits probabilities of 0 and 1. The first probabilistic model was Lord's two-parameter normal ogive model in which the item characteristic curve is in the form of a cumulative normal function. The function is characterized by two parameters: an item difficulty parameter and a discrimination parameter. Birnbaum replaced the normal ogive with the two-parameter logistic function. The functions are similar in form, but the logistic function has some technical advantages. The logistic item characteristic curve also has an item difficulty and discrimination parameter. The Rasch model specifies a logistic function as well, but a one-parameter function. The discrimination parameters of the test items are set equal to each other, and the item characteristic curve is determined by only one parameter (i.e., item difficulty). A model for multiple-choice items is Birnbaum's three-parameter model. The item characteristic curve is also a logistic function. The curve is determined not only by an item difficulty and discrimination parameter but by an item guessing parameter. Other item response models for dichotomous items are the polynomial model (Lazarsfeld & Henry, 1968, chap. 7) and the Mokken (1970) model. In the polynomial model, the item characteristic curve is a polynomial function; in the Mokken model, the curve is a nonparametric, monotone increasing function of the latent trait.

The models just mentioned are not based on any assumptions about the distribution of the latent trait(s). They can be extended with the assumption of a latent trait distribution. For example, it can be assumed that the latent trait has a normal distribution (or the latent traits have a multivariate normal distribution) in a population of subjects.

It is assumed that subjects' item response behavior is determined by one or more latent traits. A latent trait is considered to be an unobserved continuous variable. The assumption that the latent variable is an unobserved nominal variable gives rise to latent class models. Examples are the state models for mastery testing in educational measurement (Macready & Dayton, 1980). In these models, two latent classes of masters and nonmasters are assumed. Masters are examinees who have mastered the contents of the test. Although these examinees have mastered the contents, they can make a mistake and provide an incorrect answer to an item. Nonmasters are examinees who have not mastered the contents of the test but who can be lucky and provide a correct answer.

Item response models for dichotomous items were extended to items with more than two answer categories (polytomous items). The models are the same type used for dichotomous items, but they have been modified to make them appropriate for multicategory answer scales. For example, the assumption of the point binomial distribution of dichotomous items is replaced by the assumption of a multinomial distribution. It is assumed that a subject chooses exactly one of a number of answer categories. The sum of the probabilities of the subject choosing each of the categories is equal to 1.

Generalized Linear Item Response Theory

In educational and psychological measurement, subjects respond to items on tests, questionnaires, and inventories. The items differ in format and require different types of responses. These tests can consist of questions that can be answered correctly or incorrectly (dichotomous items), the choice of one option out of a number of graded options (Likert scale), marking a point on a continuous line, and the time needed to complete an item.

This article presents assumptions that are generalizations or modifications of the assumptions of IRT models for dichotomous items. The first assumption of IRT for dichotomous items is that the distribution of the subject's item responses is a point binomial distribution. In GLIRT, the item format is not restricted to dichotomous items, which means that a variety of item formats can be used. For each of the item formats, an appropriate item response distribution must be assumed. For example, the normal distribution can be assumed for continuous item responses and the multinomial distribution can be assumed for responses to polytomous items. The second assumption of IRT is that the item responses are explained by one or more latent traits. In GLIRT, this assumption is generalized as follows: One latent variable, a continuous latent variable (latent trait) or a nominal latent variable (latent class), is assumed. In addition to this latent variable, p other variables can be used to explain a subject's response behavior. Each of these p variables can be latent or observed, as well as continuous or nominal; that is, each can be of the observed–continuous, observed–nominal, latent–continuous (latent trait), or latent–nominal (latent class) type. The third assumption of IRT is local independence. In GLIRT, the same assumption of local independence is made. The fourth assumption of IRT involves the regression of the item responses on the latent trait(s). The regression function of IRT relates the expected item response (i.e., for dichotomous items, the probability of the subject providing a correct response) to the latent trait(s). In GLIRT, the regression function relates the transformed expected item response to the latent variable and the other p explanatory variables. The observed response of subject i to test item j has the expected value τ_{ij}, where the expectation is defined over (hypothetical) repeated administrations of the same item to the same subject or over the subpopulation of subjects with identical values on the explanatory (latent and observed) variables. It is assumed, as in generalized linear models (McCullagh & Nelder, 1989), that there is a monotone differentiable function g of the expected item response such that a linear combination of latent and observed explanatory variables is a predictor of $g(\tau)$; that is,

$$g(\tau_{ij}) = b_j + a_j t_i + c_{1j} z_{1i} + \cdots + c_{pj} z_{pi}, \qquad (1)$$

where b_j, a_j, and $c_{1j}, \ldots, c_{pj}$ are parameters of item j, and t_i and $z_{1i}, \ldots, z_{pi}$ are subject i's scores on $(p + 1)$ variables. The z variables may be observed or latent variables; t is a latent variable that is used in all item response models (and is therefore denoted by t instead of z). Moreover, t and z may be discrete nominal or continuous variables. The origin and scale of continuous latent variables are determined by setting their mean at 0 and variance at 1 and by assuming that the function g is monotonically increasing in continuous latent variables. Nominal variables are handled by using dummy coding (e.g., for a dichotomous nominal variable such as sex, one category is coded as 0 and the other as 1; for nominal variables with more than two categories, a set of dummy variables is used).

Formula 1 is the model for the regression of the observed item responses on fixed values of nominal or continuous latent or observed explanatory variables. The structure of Formula 1 is linear. In general, however, the regression function itself is nonlinear because of the transformation function $g(\tau_{ij})$.

The theory can be extended further by assuming a distribution for the latent variables in a population of subjects (e.g., by assuming that continuous latent variables have a normal distribution or that nominal latent variables have a multinomial distribution).

Item Response Models

Specific item response models are derived from GLIRT in the following steps. Test items differ in format. The first step is to choose an appropriate item response distribution. The expected value of subject i's distribution of responses to item j is τ_{ij}. The second step is the selection of a transformation function $g(\tau_{ij})$ that is characteristic of the assumed item response distribution. Finally, the linear structure of Formula 1 is specified.

The generalized theory is applied to different item formats. The following formats are considered: (a) dichotomous item response, (b) continuous item response, (c) ungraded options, (d) graded options, and (e) time needed to complete the item. The generalized theory might also be applied to other item formats.

In achievement and performance testing, items are usually scored dichotomously (correct or incorrect). It is usually assumed that subject i's responses to item j have a point binomial distribution; the expected value τ_{ij} is the probability of subject i having the correct answer to the jth item. The probability is between 0 and 1 and must be transformed to a scale from minus to plus infinity. Appropriate transformation functions for the binomial distribution are the logit, probit, and complementary log–log functions (McCullagh & Nelder, 1989, Section 2.2.3). Finally, specifications of the variables and parameters of Formula 1 are made.

The procedure is applied to dichotomously scored items to demonstrate the derivation of specific models. Suppose it is assumed that the item responses are binomially distributed with expected value τ_{ij}. An appropriate transformation function is the logit function. If it is specified in Formula 1 that t is a continuous latent trait and $c_{1j} = c_{2j} = \cdots = c_{pj} = 0$, it follows that

$$g(\tau_{ij}) = \ln\{\tau_{ij}/(1 - \tau_{ij})\} = b_j + a_j t_i = a_j(\theta_i - b_j^*), \qquad (2)$$

where $\theta_i = t_i$ and $b_j^* = -b_j/a_j$. From Formula 2, it follows that the nonlinear regression of the observed item response on the latent trait is

$$\tau_{ij} = \exp\{a_j(\theta_i - b_j^*)\}/[1 + \exp\{a_j(\theta_i - b_j^*)\}], \qquad (3)$$

which is the item characteristic curve of Birnbaum's two-parameter model (Hambleton & Swaminathan, 1985, Section 3.3.2). The model has an item difficulty (b_j^*) and an item discrimination (a_j) parameter. If it is specified that all items have identical discrimination parameters, the result is the one-parameter Rasch model (see Table 1). If it is further specified that all items have both identical discrimination and difficulty parameters, the result yielded is the binomial error model (Lord & Novick, 1968, chap. 23). The logit transformation was applied to derive these models. If the inverse cumulative normal distribution (probit) transformation (McCullagh & Nelder, 1989, Section 2.2.3) is used in combination with the same specification of Formula 1, as for Birnbaum's model, the result is the two-parameter normal ogive model (Lord & Novick, 1968, Section 16.5). In these models, it is assumed that the latent variable t is a continuous trait. Other models are obtained by assuming that the latent variable is nominal. For example, if Formula 2 is used and if it is assumed that t is a nominal variable with two latent classes ($t_i = 0$ for the first class and $t_i = 1$ for the second class), the result is Macready and Dayton's (1980) state model for mastery testing (see Table 1). This model makes a distinction between latent classes of masters and nonmasters. As mentioned earlier, masters ($t_i = 1$) are subjects who have mastered the content of an achievement test item. They have a probability of providing a correct answer to the item as well as a probability of providing an incorrect answer, which is called the omission error. Nonmasters ($t_i = 0$) have not mastered the content of the item, but they also have a probability of providing the correct answer, which is called the intrusion error.

The second item format that is considered consists of a continuous scale (e.g., marking a point on a continuous line). An approximation is a Likert scale with a large number of categories. It is assumed that subject i's responses to item j are normally distributed with homogeneous variance. The transformation function for the normal distribution is the identity function (McCullagh & Nelder, 1989, Section 2.2.4). Using this transformation function and making specifications in Formula 1 yield, among others, Jöreskog's (1971) factor-analytic model for congeneric item responses (see Table 1). The third item format consists of ungraded options, as in questionnaires; the subject chooses one of the options. It is assumed that the item responses have a multinomial distribution over the options. The transformation function is the logit of one arbitrarily selected option and each of the other options. One specification of Formula 1 yields Bock's (1972) model for nominal options (see Table 1). The fourth item format consists of graded options such as Likert-scale items. It is assumed that the item responses have a multinomial distribution over the options. The ordinal nature of the options is retained by using, for example, cumulative or adjacent-category logit transformation functions (Agresti, 1990, chap. 9). The cumulative proportions logit in combination with one Formula 1 specification yields Samejima's (1969) model, whereas adjacent-ratio logits in combination with a different Formula 1 specification yield Masters's (1982) partial credit model (see Table 1). The final item format considered is the time needed to complete an item. Suppose that the time needed to complete an item of a speed test is exponentially distributed. The expected item response is the mean time to complete the item. The transformation function is the inverse function $1/\tau_{ij}$ (McCullagh & Nelder, 1989, Section 2.2.4). One Formula 1 specification yields Rasch's (1960, chap. 3) model for reading speed (see Table 1).

An overview of specific models is provided in Table 1, which contains (a) the item format, (b) the assumed item response distribution, (c) the transformation function, and (d) the specification of Formula 1. The table shows that GLIRT subsumes as special cases most of the item response models described in the psychometric literature.

Table 1 contains probabilistic item response models. The Guttman model, a deterministic model, is not included in the table but can be described as an extreme case of the generalized theory. It is the limiting case of the one-parameter Rasch or normal ogive model in which the slope parameter a goes to plus infinity (Thissen & Steinberg, 1986, p. 569).

The Mokken (1970) model cannot be subsumed under GLIRT. The model has a nonparametric item characteristic

Table 1
Overview of Item Response Models Subsumed Under Generalized Linear Item Response Theory

Item format	Item response distribution	Transformation $[g(\tau_{ij})]$	Specification of Formula 1	Model
Dichotomous	Binomial	Logit: $\ln\{\tau_{ij}/(1 - \tau_{ij})\}$	$b_j + a_jt_i$	Birnbaum's two-parameter model
			$b_j + at_i$	Rasch's one-parameter model
			$b + at_i$	Binomial error model
			b_j for $t_i = 0$; $b_j + a_j$ for $t_i = 1$	Macready and Dayton's (1980) state model for mastery testing
			$b_j + a_jt_i + c_{1j}t_{1i} + \cdots + c_{pj}t_{pi}$	Reckase's (1985) multidimensional model with $(p + 1)$ latent traits
		Cumulative normal (probit): $\Phi^{-1}(\tau_{ij})$	$b_j + a_jt_i$	Lord's normal ogive model
		Complementary log–log: $\ln\{-\ln(1 - \tau_{ij})\}$	$b_j + at_i$	Goldstein's (1980) model
Continuous	Normal with homogeneous item response variance σ_j^2	Identity: τ_{ij}	$b_j + a_jt_i$	Jöreskog's (1971) model for congeneric item responses
			$b + at_i$, $\sigma_j^2 = \sigma^2$	Lord and Novick's (1968) model for parallel item responses
			$b + at_i$	Lord and Novick's (1968) model for tau-equivalent item responses
			$b_j + at_i$	Lord and Novick's (1968) model for essentially tau-equivalent item responses
			$b_j + a_jt_i + c_{1j}t_i^2 + \cdots + c_{pj}t_i^{p+1}$	McDonald's (1982) polynomial model
		τ_{ikl}	$b_{kl} + a_{kl}t_i + c_{1kl}t_{ki} + c_{2kl}t_{li} + c_{3kl}t_{kli}$	Mellenbergh, Kelderman, Stijlen, and Zondag's (1979) model for items from the combination of the *k*th element of the first facet and the *l*th element of the second facet of a facet design
Ungraded options	Multinomial	Logit of *k*th category and fixed *m*th category: $\ln(\tau_{ijk}/\tau_{ijm})$	$b_{jk} + a_{jk}t_i$	Bock's (1972) model for nominal options
Graded options	Multinomial	Cumulative logit of *k*th category: $\ln\{(\tau_{ijk} + \cdots + \tau_{ijm})/(\tau_{ij1} + \cdots + \tau_{ijk-1})\}$	$b_{jk} + a_jt_i$	Samejima's (1969) model for cumulative proportions
		Adjacent *k*th and $(k - 1)$th category logit: $\ln(\tau_{ijk}/\tau_{ijk-1})$	$b_{jk} + at_i$	Masters's (1982) partial credit model
			$\delta_i + \gamma_k + at_i$	Andrich's (1978) rating scale model with location (δ_i) and threshold (γ_k) parameters
Response time	Exponential	Inverse: $1/\tau_{ij}$	$b_j + at_i$	Rasch's (1960) model for reading speed

Note. The symbol t is used for latent variables; τ_{ijk} denotes subject *i*'s probability of choosing the *k*th option of item *j* for items with more than two options; τ_{ikl} denotes subject *i*'s mean response to the item from the combination of the *k*th element of the first facet and the *l*th element of the second facet for continuous responses.

curve that cannot be described as a special case of the parametric function of Formula 1. Other examples of models that cannot be subsumed under GLIRT are Birnbaum's three-parameter logistic model (Lord & Novick, 1968, Section 17.3) and Thissen and Steinberg's (1984) model for multiple-choice items. According to these models, the probability of a correct answer is determined not only by the probability of the subject knowing the correct answer but by the probability of the subject guessing the correct answer. Thissen and Steinberg (1986) referred to these models as left-side added because the probability of guessing is added at the left (i.e., low) side of the latent variable. Subjects at low values of the latent variable do not know the correct answer but can obtain the correct answer by guessing. Generally speaking, the sum of the probability of guessing and the probability of knowing the correct answer for nonguessing subjects cannot be transformed to a linear form by applying the usual transformation functions.

The overview of specific item response models provided in Table 1 demonstrates the following points. First, for some item response distributions, different transformation functions are appropriate, and they yield different item response models (e.g., the logit transformation for dichotomous, binomially distributed item responses yields the Birnbaum model, whereas the probit transformation yields the normal ogive model). Second, the specification of continuous latent variables leads to latent trait models whereas the specification of nominal latent variables leads to latent class models. Third, the specification of the parameters of Formula 1 also leads to different models. Fourth, the various specifications of Formula 1 are not restricted to one item format but can be applied to all other item formats.

Psychometric Concepts

Many psychometric concepts were developed within the context of specific item response models such as Birnbaum's two-parameter model and the Rasch model for dichotomous, binomially distributed item responses. Most of the specific models represent special cases of the generalized theory. Some concepts that were developed for specific models might be useful for all members of the GLIRT family. In this section, the concepts of differential item functioning, specific objectivity, information, and reliability are subsumed under the generalized theory.

Differential Item Functioning

The concept of item bias or differential item functioning was developed within the context of item response theory for dichotomous, binomially distributed item responses. Formula 3 is the item characteristic curve of Birnbaum's two-parameter model. In this model, an item is considered to be biased between two groups when the item characteristic curves of the two groups do not coincide (e.g., see, Mellenbergh, 1989).

Differential item functioning is subsumed under the generalized theory as follows. Formula 1 is specified to contain three explanatory variables:

$$g(\tau_{ij}) = b_j + a_j t_i + c_{1j} z_{1i} + c_{2j} z_{2i}. \qquad (4)$$

The variable t is the latent variable measured by the item; it is called the measuring variable. The variable z_1 is supposed to cause the bias; it is called the biasing variable. The variable z_2 is the interaction of the measuring and biasing variables; the interaction of two variables is obtained by using their product. Therefore, the general model for differential item functioning is

$$g(\tau_{ij}) = b_j + a_j t_i + c_{1j} z_{1i} + c_{2j} t_i z_{1i}. \qquad (5)$$

The biasing variable is denoted by z to indicate that it may be an observed or latent variable. As stated earlier, both the measuring variable and the biasing variable may be continuous or nominal. Examples of observed nominal and continuous biasing variables are sex and age, respectively. An example of a latent nominal biasing variable is the variable consisting of the latent classes of providing and not providing socially desirable answers; an example of a continuous latent biasing variable is the latent trait of social desirability.

Two types of differential item functioning are distinguished. Formula 5 contains not only the effect of the measuring and biasing variables but their interaction. Therefore, Formula 5 is the model for nonuniform differential item functioning. If the model does not contain the interaction term (i.e., $c_{2j} = 0$), it is the model for uniform differential item functioning.

The concept of differential item functioning in Birnbaum's two-parameter model is a special case of the Formula 5 model. The items are dichotomously scored, and the item responses are binomially distributed. Suppose that item j functions differentially between a majority and a minority group. The biasing variable is a nominal group membership variable that is coded 0 for members of the majority group and 1 for members of the minority group. From the logit transformation and the differential item functioning model Formula 5, it follows that

$$g(\tau_{ij}) = \ln\{\tau_{ij}/(1 - \tau_{ij})\} = b_j + a_j t_i, \qquad (6a)$$

for members of the majority group ($z_{1i} = 0$), and

$$g(\tau_{ij}) = \ln\{\tau_{ij}/(1 - \tau_{ij})\} = b_j + a_j t_i + c_{1j} + c_{2j} t_i = b_j^* + a_j^* t_i, \qquad (6b)$$

where $b_j^* = b_j + c_{1j}$ and $a_j^* = a_j + c_{2j}$, for members of the minority group ($z_{1i} = 1$). A comparison of Formulas 6(a) and 6(b) shows that, in general, the item characteristic curves of the majority and minority group do not coincide, which is the definition of differential item functioning under Birnbaum's two-parameter model for dichotomous, binomially distributed item responses.

The concept of differential item functioning is very general. The Formula 5 model can be applied to different item formats and item response distributions, and it includes (a) a latent trait or a latent class measuring variable, (b) a continuous or a nominal biasing variable, (c) an observed or latent biasing variable, and (d) uniform and nonuniform differential item functioning.

Specific Objectivity

The concept of specific objectivity was introduced within the context of the Rasch model for dichotomous item responses. A measurement is specifically objective if the comparison of two subjects' properties does not depend on the particular subset of items used and the comparison of two items' properties does not depend on the particular subgroup of subjects tested (Fischer, 1987, p. 567).

The concept of specific objectivity is subsumed under the generalized theory as follows. A measurement is specifically objective if two conditions are fulfilled: (a) For the comparison of two subjects, i and i', the difference $g(\tau_{ij}) - g(\tau_{i'j})$ does not depend on item j, and (b) for the comparison of two items, j and j', the difference $g(\tau_{ij}) - g(\tau_{ij'})$ does not depend on subject i. These conditions are fulfilled if Formula 1 is specified by assuming that the a and c parameters do not depend on the item; that is,

$$g(\tau_{ij}) = b_j + a t_i + c_1 z_{1i} + \cdots + c_p z_{pi}. \qquad (7)$$

The comparison of subjects i and i' is the difference

$$g(\tau_{ij}) - g(\tau_{i'j}) = a(t_i - t_{i'}) + c_1(z_{1i} - z_{1i'}) + \cdots + c_p(z_{pi} - z_{pi'}), \qquad (8a)$$

which does not depend on item j. The comparison of items j and j' is the difference

$$g(\tau_{ij}) - g(\tau_{ij'}) = b_j - b_{j'}, \qquad (8b)$$

which does not depend on subject i.

The generalized concept of specific objectivity subsumes the concept that was developed within the context of the Rasch model. The Rasch model is obtained by specifying $a_j = a$ in Formula 2 (see Table 1). Therefore, the Rasch model is the special case of Formula 7 where $c_1 = c_2 = \cdots = c_p = 0$. Formula 7 shows that the concept of specific objectivity is generalized. The number of explanatory variables may be larger than one, the latent variable t may be continuous or nominal, and the z variables may be continuous or nominal and observed or latent.

Reliability and Information

The concept of reliability was developed within the context of classical test theory. According to classical test theory, the observed score is usually used as an estimate of the true score. The reliability of this estimate is defined as the squared correlation of the observed and true scores in a population of subjects (Lord & Novick, 1968, Section 3.4). In specific models of GLIRT, different methods are used to estimate subject i's latent value t_i; an estimate of t_i is denoted by $\hat{t}_i$. Analogous to classical test theory, the reliability of an estimate of a continuous latent trait is defined as the squared correlation of the latent trait and its estimate:

$$\{\mathrm{cor}\,(t, \hat{t})\}^2 = \{\mathrm{cov}\,(t, \hat{t})\}^2/\{\mathrm{var}\,(t)\ \mathrm{var}\,(\hat{t})\}, \qquad (9)$$

where cor, cov, and var denote correlation, covariance, and variance, respectively. This definition is rather general because it can be applied to all latent trait models of GLIRT. Moreover, the reliability is defined for different estimates of the latent trait by using different estimation methods, (i.e., each estimate has its own reliability). The concept of reliability is not used for most of the item response models of Table 1. However, the concept can be defined for latent trait models and can be of use in applications. An example is the reliability of the maximum likelihood estimate of the latent trait in Jöreskog's (1971) congeneric model for continuous, normally distributed item responses, which is useful in the area of test construction (Mellenbergh, 1993).

The concept of information was used in the context of models for dichotomous, binomially distributed item responses. It is defined as the inverse of the variance of the maximum likelihood estimate $\hat{t}_i$ at subject i's latent trait value t_i (Hambleton & Swaminathan, 1985, Section 5.4):

$$I(t_i) = 1/\mathrm{var}\,(\hat{t}_i\,|\,t_i). \qquad (10)$$

The concept of information is not restricted to models for dichotomous, binomially distributed item responses but applies to all models of GLIRT. An example is the information function of the maximum likelihood estimate in Jöreskog's (1971) congeneric model for continuous, normally distributed item responses (Mellenbergh, 1993).

The concepts of reliability and information apply to all latent trait models of GLIRT, but they have different interpretations. Reliability is a population-dependent global concept of measurement precision. As shown by Formula 9, it is the proportion of latent trait variance that is linearly predicted by the latent trait estimate in a population of subjects. On the other hand, information is a population-independent local concept of measurement precision. As shown by Formula 10, it is monotonically related to the variance of the latent trait estimate $\hat{t}_i$ at the latent trait level t_i. This variance is defined as the variance of the latent trait estimate in repeated (hypothetical) administrations of the same test to the same subject or in the subpopulation of subjects with identical latent trait values.

Discussion

This article has presented a description of a general theory for item responses. The theory makes the following assumptions: (a) a distribution of the responses to a given item format, (b) explanation of the item responses by one continuous or nominal latent variable and p latent and observed variables that are continuous or nominal, (c) local independence of responses to different test items given the values on the explanatory variables, and (d) existence of a monotone differentiable function g of the expected item response τ such that a linear combination of the explanatory variables is a predictor of $g(\tau)$. The theory can be further extended by assuming a distribution of the latent explanatory variables. The theory subsumes as special cases most of the item response models described in the psychometric literature.

GLIRT has some advantages over specific item response models. First, concepts that have been used for only one type of item response model might be applied to all item response models that are covered by the generalized theory. Examples are the concepts of specific objectivity and differential item functioning. These concepts were developed in the context of models for dichotomous items but can be applied to a much larger set of item response models if used in the framework of the generalized theory. Second, models developed within the context of one type of item format can be applied to all item formats. Macready and Dayton's (1980) state model for mastery testing, a latent class model for dichotomous, binomially distributed item responses, is one such example. Latent class models can also be applied to other item formats by using a discrete latent variable and dummy coding. Third, the generalized theory was applied to five types of item formats. It is possible that the generalized theory also applies to other item formats. The theory can be used when it is possible to assume a suitable distribution of the item responses and find an appropriate link function.

According to the statistical theory of generalized linear models, as described by McCullagh and Nelder (1989), all explanatory variables in Formula 1 are observed variables. One general algorithm and one computer program (GLIM) are used for testing models and estimating parameters. In the case of latent explanatory variables, only specific algorithms and programs were developed, such as BILOG (Mislevy & Bock, 1983) for the Birnbaum model for dichotomous item responses and LISREL (Jöreskog & Sörbom, 1988) for the factor-analytic models for continuous item responses. It would be worthwhile to develop a general algorithm and program for mixed latent and observed explanatory variables. In developing such an algorithm, it might be useful to assume a distribution of the latent variables and to apply the EM algorithm (e.g., see Little & Rubin, 1987, chap. 7).

In short, most of the known psychometric models and concepts can be subsumed under the generalized theory. The generalized theory applies to responses to items of different formats, which implies that the term *item response theory* should not be restricted to theories for dichotomous items. The latent variable can be nominal, which means that the term *latent trait theory* is also too restricted. It is for this reason that the term *generalized linear item response theory* has been proposed.

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

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

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

Dobson, A. J. (1983). *An introduction to statistical modelling.* London: Chapman & Hall.

Fischer, G. H. (1987). Applying the principles of specific objectivity and of generalizability to the measurement of change. *Psychometrika, 52,* 565–587.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33,* 234–246.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42,* 139–167.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL VII: A guide to the program and applications.* Chicago: SPSS, Inc.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Macready, G. B., & Dayton, C. M. (1980). The nature and use of state mastery models. *Applied Psychological Measurement, 4,* 493–516.

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

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49,* 529–544.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6,* 379–396.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127–143.

Mellenbergh, G. J. (1993). The unidimensional latent trait model for continuous item responses in brief. In R. Steyer, K. F. Wender, & K. F. Widaman (Eds.), *Psychometric Methodology, Proceedings of the 7th European Meeting of the Psychometric Society in Trier* (pp. 333–337). New York: Gustav Fischer Verlag.

Mellenbergh, G. J., Kelderman, H., Stijlen, J. G., & Zondag, E. (1979). Linear models for the analysis and construction of instruments in a facet design. *Psychological Bulletin, 86,* 766–776.

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software.

Mokken, R. J. (1970). *A theory and procedure of scale analysis.* The Hague, The Netherlands: Mouton.

Moosbrugger, H., & Müller, H. (1982). A classical latent additive test model. *German Journal of Psychology, 6,* 145–149.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks paedagogiske Institut.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs, 17* (4, Pt. 2).

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49,* 501–519.

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

Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement, 7,* 211–226.

Received February 14, 1992
Revision received May 12, 1993
Accepted May 21, 1993 ■