A Conditional Item-Fit Index for Rasch Models

Jürgen Rost and Matthias von Davier
Institute for Science Education—IPN

A new item-fit index is proposed that is both a descriptive measure of deviance of single items and an index for statistical inference. This index is based on the assumptions of the dichotomous and polytomous Rasch models for items with ordered categories and, in particular, is a standardization of the conditional likelihood of the item pattern that does not depend on the item parameters. This approach is compared with other methods for determining item fit. In contrast to many other item-fit indexes, this index is not based on response-score residuals. Results of a simulation study illustrating the performance of the index are provided. An asymptotically normally distributed Z statistic is derived and an empirical example demonstrates the sensitivity of the index with respect to item and person heterogeneity. *Index terms: appropriateness measurement, item discrimination, item fit, partial credit model, Rasch model.*

The assessment of item fit and person fit for item response models and, in particular, for Rasch models has somewhat different traditions in mental test theory. Item fit often has been treated as part of global model fit; for example, the conditional likelihood ratio test by Andersen (1973) that compares the item parameter estimates for different score groups. The Martin-Löf test (Martin-Löf, 1973) and the Q_1 statistic (van den Wollenberg, 1979, 1982) also test the constancy of item parameters in different score groups, even though these tests do not require separate estimates within each score group.

In addition to this tradition of summarizing item-specific deviations of observed and expected response frequencies as global fit statistics (for an overview see Glas, 1989; van den Wollenberg 1988), there also has been some use of item-fit statistics for evaluating and selecting single items. These statistics are discussed below.

Research on person-fit measures has flourished during the past decade. Person-fit measures also have been referred to as appropriateness measures, caution indexes, and measures that detect aberrant response patterns (Drasgow, Levine, & Williams, 1985; Levine & Drasgow, 1982; Molenaar & Hoijtink, 1990; Tatsuoka & Linn, 1983). Person-fit research is important because decisions based on test scores are made about individuals. However, test construction is concerned primarily with problems of item evaluation and selection. Yet there is a strong symmetry between item-fit and person-fit measures; therefore, properties of item- and person-fit indexes can be analyzed simultaneously (Reise, 1990).

Research on item- and person-fit measures usually is based on item response theory (IRT), and most indexes can be applied to any IRT model and even to non-IRT models. This is useful because the fit of different models may be compared directly using these measures. From a theoretical point of view, it is interesting that Rasch measurement theory has not developed its own measures of person- and item-fit. The properties of Rasch models—parameter separability and conditional inference—allow for theoretically elegant ways of assessing model fit. Therefore, it seems appropriate that these properties should be used for the evaluation of single items and persons.

An item-fit index is derived below that is symmetric to a person-fit index developed earlier (Tarnai & Rost, 1990). The properties of this index, named "item-Q" (in contrast to the "person-Q" index by Tarnai & Rost),

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 2, June 1994, pp. 171–182

0146-6216/94/020171-12$1.85

are shown using simulated and real data. In order to summarize the properties of the item-Q index, a distinction has to be made between item-Q as a descriptive measure used to compare the fit of two or more items of a test and as a measure used to test the significance of the fit of single items to the Rasch model. As a descriptive measure of deviance, item-Q is standardized. It ranges from 0 to 1 with a midpoint of .5, which indicates random response behavior. A 0 value represents the most likely response vector, and a value of 1 the least likely response vector for an item. For descriptive purposes, the index does not require an estimate of the item parameter(s) for the item under consideration but is conditioned on the score distribution of that item.

When using item-Q to test the significance of the fit of single items, however, the item parameters must be estimated beforehand and are used to derive the sampling distribution. Item-Q is not based on differences between observed and expected response scores; therefore, it does not deal with problems arising from the discrete nature of response scores. Furthermore, it can be applied to any unidimensional Rasch model, such as the dichotomous model, the rating scale model (Andrich, 1978), the dispersion model (Andrich, 1982), the partial credit model (Masters, 1982) or the successive interval model (Rost, 1988).

Measures of Item Fit

Measures of item fit can be divided into three groups of indexes:

1. χ^2 statistics that compare observed and expected response frequencies in a priori defined groups of examinees;
2. Standardized Z values based on the likelihood function of the response or item pattern; and
3. Measures based on score residuals, that is, based on averaged deviations of observed and expected item responses.

All three approaches are described briefly and evaluated with regard to how they fit with the assumptions and statistical properties of Rasch models.

The χ^2 Approach

The χ^2 approach was proposed by Wright & Panchapakesan (1969) and was discussed by Bock (1972) and van den Wollenberg (1979). It is primarily used to assess whether the postulated shape of the item response function (IRF) holds for the observed data on a particular item. Persons are grouped using their test scores or trait level estimates ($\hat{\beta}$). Let n_j denote the number of individuals in β level or group $j (j = 1, \ldots, J)$. Then the χ^2 statistic for item i is defined as:

$$\chi_i^2 = \sum_{j=1}^{J} \frac{n_j (o_{ij} - e_{ij})^2}{e_{ij}(1 - e_{ij})}, \tag{1}$$

where o_{ij} and e_{ij} are the observed and expected proportions of correct (1) responses to item i in group j, respectively. Yen (1981), McKinley & Mills (1985), and Reise (1990) studied the performance of fit statistics based on this approach.

The χ^2 approach is not restricted to a particular IRT model. The calculation of χ_i^2 requires estimates of β parameters or a mean estimate for all individuals in a group. Item parameter estimates are required to calculate the expected proportion of correct responses for each group.

Although statistical tests based on grouping data represent a basic principle of testing statistical models, their power to detect misfits clearly depends on whether the grouping selected reflects the type of misfit in the data. If, for example, an item has an observed IRF that deviates from the IRF postulated by the model, grouping according to the scores of the individuals may reveal this misfit. If there are two or more types of individuals at each β level, and a different item difficulty holds for each, this is not necessarily revealed by score grouping.

Therefore, it is necessary to have specific hypotheses about possible misfit in order to form relevant

groups. It even may be that a variable splitting the sample in such a way that item misfit becomes apparent has not been observed or is not observable at all; for example, when different people use different solution strategies to the same tasks and, hence, different item parameters hold for these people. This is the case in which discrete mixture distribution models, such as the mixed Rasch model (Rost, 1990), can be applied successfully. In these models, latent classes of individuals are identified so that different sets of model parameters apply in each latent population. However, this approach to testing the Rasch model has only been developed for sets of items and not for testing single items (Rost & Von Davier, 1992). A disadvantage of the χ^2 approach for testing item fit is that it cannot be generalized easily to polytomous ordinal IRT models because it is frequency-based and additional assumptions must be introduced to handle frequencies of ordered categories.

The Likelihood-Based Approach

The likelihood-based approach was proposed by Levine & Rubin (1979) and Drasgow et al. (1985) for studying person fit. Reise (1990) investigated the related Z statistic for evaluating item fit. The likelihood L_i of an (dichotomous) item pattern for a person is defined as

$$L_i = \prod_{v=1}^{N} p_{vi}^{x_{vi}} (1 - p_{vi})^{1-x_{vi}}, \tag{2}$$

where x_{vi} is the (0,1) response of individual v to item i, and p_{vi} is the response probability of person v to item i as defined by any IRT model. Because L_i strongly depends on the difficulty of the items (or the β level in the case of person fit), Drasgow et al. (1985) proposed a standardization of the likelihood which makes use of the fact that maximum-likelihood estimators are normally distributed. It follows that

$$\log L_{vi} = x_{vi} \log p_{vi} + (1 - x_{vi}) \log(1 - p_{vi}) \tag{3}$$

which has, under model assumptions, an expected value

$$\mathrm{E}_{vi} = p_{vi} \log p_{vi} + (1 - p_{vi}) \log(1 - p_{vi}) \tag{4}$$

and variance

$$\mathrm{V}_{vi} = p_{vi}(1 - p_{vi})\left[\log p_{vi} - \log(1 - p_{vi})\right]^2. \tag{5}$$

Hence, the fit statistic is based on Z values,

$$Z_{vi} = \frac{\log L_{vi} - \mathrm{E}_{vi}}{(\mathrm{V}_{vi})^{1/2}}, \tag{6}$$

and can either be accumulated over persons for single items (for studying item fit) or over items for single persons (for studying person fit) (Reise, 1990). These sums, then, are asymptotically normally distributed.

This approach can be demonstrated easily. Suppose the response probability of person v to item i is .9 for a correct response and .1 for an incorrect response. The maximum of the log-likelihood (i.e., the log-likelihood of a correct response) is $\log(.9) = -.105$, and its minimum is $\log(.1) = -2.3$ (for an incorrect response). The expectation according to Equation 4 is $-.325$, which splits the interval between the minimum and the maximum according to the proportion of the given response probabilities (9:1). Therefore, positive Z values

are obtained when the most probable response is given more often than expected and negative values result for inconsistent responses (Reise, 1990). This likelihood-based approach for assessing item fit is, like the χ^2 statistic, applicable to any IRT model and requires, even in the case of the Rasch model, the estimation of both item and person parameters.

The Score Residual Approach

Item and person parameters also must be estimated for the score residual approach, which was developed primarily within Rasch measurement theory (Rasch, 1980; Wright, 1980). In this approach, item fit is evaluated on the basis of the deviation of observed and expected item responses (i.e., on score residuals).

Again, let p_{vi} denote the probability of a correct response under some model and x_{vi} the (0,1) response to the item. The mean square statistic based on standardized residuals is

$$U_{vi} = \frac{x_{vi} - p_{vi}}{\left[p_{vi}(1-p_{vi})\right]^{1/2}} . \tag{7}$$

This can be formed by summing squared residuals over persons (item fit) or over items (person fit) (Wright, 1980; Wright & Stone, 1979). These mean squares may be transformed into t statistics that approximate a unit normal distribution. This approach can be generalized easily to polytomous ordinal item responses (Wright & Masters, 1982), where it may be even more appropriate because the expected score can be compared with more than two observed scores.

The Item-Q Index

The item-Q index has a different rationale. It uses Rasch measurement principles--the item parameter is conditioned out of the item-fit index. Item-Q is based on the likelihood of observed response patterns (as is the person-fit approach), but it uses conditional likelihoods (i.e., the likelihood of an item pattern conditional on the item score). Thus, an item-fit index is obtained that, in some sense, is "parameter free" with respect to the item parameter. However, for statistical inference with this index an estimate of the item parameter is required.

The item-Q index is derived for the ordinal Rasch model [called the partial credit model by Masters (1982)], which is a generalization of the dichotomous Rasch model and various ordinal models (Andrich, 1978, 1982; Rost, 1988). The model describes the response probability of category x ($x = 0, 1, \ldots, m$) as a logistic function of β_v and item-category parameters α_{ix}, which can be interpreted as cumulated threshold parameters μ_{ix} (Andrich, 1978; i.e., $\alpha_{ix} = \sum_s \mu_{is}$ and $\alpha_{io} = 0$),

$$p(X_{vi} = x) = \frac{\exp(x\beta_v + \alpha_{ix})}{\sum_s \exp(s\beta_v + \alpha_{is})} . \tag{8}$$

The fit of item i is evaluated with regard to its observed item response vector $\mathbf{x}_i$, which is an element of the space $\Omega_{\mathbf{x}}$ of all possible item vectors of length N and with components $x_v \in \{0, 1, \ldots, m\}$

$$\Omega_{\mathbf{x}} := \left\{\mathbf{x} = (x_1, x_2, \ldots, x_n), x_v \in (0, 1, \ldots, m)\right\} . \tag{9}$$

However, for evaluating an item, the probability distribution of the entire vector space $\Omega_{\mathbf{x}}$ is not required, but rather the subspace of $\Omega_{\mathbf{x}}$ that covers all vectors with given item score frequencies is required. Let n_{ix} denote the frequency of category x with pattern $\mathbf{x}$. Then the restricted vector space is

$$\Omega_{\mathbf{x}|n_{ix}} := \left\{\mathbf{x} | n_0(\mathbf{x}) = n_{i0}, \ldots, n_m(\mathbf{x}) = n_{im}\right\} . \tag{10}$$

The rationale for considering only the patterns with fixed score frequencies is that the total score of each category—which is the frequency of responses in each category of an item, n_{ix}—is a sufficient statistic for estimating the item parameters (just as the number of correct responses is for a dichotomous item). Because the items and their categories are allowed to vary in difficulty, it does not make sense to compare the probability of a given item pattern with item patterns that are associated with different item difficulties.

The probability distribution of $\Omega_{\mathbf{x}|n_{ix}}$ is defined by the conditional pattern probabilities, where n_{ix} is the number of responses in category x with item i,

$$p(\mathbf{x}_i|n_{ix}) = \frac{p(\mathbf{x}_i)}{\sum_{\mathbf{x}|n_{ix}} p(\mathbf{x}|n_{ix})} = \frac{\exp\left(\sum_v x_{vi}\beta_v + n_{ix}\alpha_{ix}\right)}{\sum_{\mathbf{x}|n_{ix}} \exp\left(\sum_v x_v\beta_v + n_{ix}\alpha_{ix}\right)} = \frac{\exp\left(\sum_v x_{vi}\beta_v\right)}{\sum_{\mathbf{x}|n_{ix}} \exp\left(\sum_v x_v\beta_v\right)}. \quad (11)$$

These do not depend on the item parameters. However, the denominator defines the symmetric functions of order $(n_{i0}, n_{i1}, \ldots, n_{im})$ of the β_vs, but these are difficult to calculate (see Rost, 1991 for an algorithm). Fortunately, they can be reduced if the ratios of two conditional likelihoods are formed. Hence, the idea of person-Q (Tarnai & Rost, 1990) and item-Q is to build a quotient (Q) of two conditional log-likelihood ratios.

The quotient quantitatively locates the likelihood of the observed pattern between the maximum and minimum likelihoods. The upper bound of the conditional pattern likelihood (maximum) is obtained for an item pattern, where the n_{im} individuals with the highest βs have responses in category m, the next $n_{i(m-1)}$ individuals in category $m-1$, and so on. This is called the Guttman pattern, $\mathbf{x}_G$, because it would be the only admissible pattern in Guttman's scalogram analysis. In Rasch models, the Guttman pattern has the highest probability of all x in $\Omega_{\mathbf{x}|n_{ix}}$.

The pattern with minimum likelihood has m responses for the n_{im} individuals with the lowest βs and a number of n_{i0} 0 responses at the other end of the β spectrum. This is called the anti-Guttman pattern, $\mathbf{x}_A$.

The likelihood ratio $\mathrm{LR}_{i,G} = p(\mathbf{x}_i|n_{ix})/p(\mathbf{x}_G|n_{ix})$ provides a partial standardization of the conditional pattern likelihood of an item i, because the ratio approximates 1 for an increasing pattern probability. However, this ratio or, in the case of log-likelihoods, this distance of observed and maximum probability

$$\log(\mathrm{LR}_{i,G}) = \log\left[p(\mathbf{x}_i|n_{ix})\right] - \log\left[p(\mathbf{x}_G|n_{ix})\right]. \quad (12)$$

must be standardized again with regard to the highest possible difference of two pattern log-likelihoods. That is,

$$\log(\mathrm{LR}_{A,G}) = \log\left[p(\mathbf{x}_A|n_{ix})\right] - \log\left[p(\mathbf{x}_G|n_{ix})\right]. \quad (13)$$

The ratio of both log-likelihood ratios varies between 0 and 1; 0 indicates perfect fit and 1 indicates perfect misfit or deviance from the model. The item-Q index thus is defined as

$$Q_i = \frac{\log(\mathrm{LR}_{i,G})}{\log(\mathrm{LR}_{A,G})}. \quad (14)$$

By inserting Equation 11, the Q_i ratio can be written as

$$Q_i = \frac{\sum_v x_{vi}\beta_v - \sum_v x_{v,G}\beta_v}{\sum_v x_{v,A}\beta_v - \sum_v x_{v,G}\beta_v} = \frac{\sum_v (x_{vi} - x_{v,G})\beta_v}{\sum_v (x_{v,A} - x_{v,G})\beta_v}, \quad (15)$$

which is a very simple function of the βs and the optimum (Guttman) and anti-Guttman responses of each person. The β parameters can be estimated using all test items, using all items except item i, or using any

other test measuring the same trait. The Guttman and anti-Guttman response of each person v (conditional on the given item score distribution) is obtained by ordering all persons according to their β level and assigning the $n_0, n_1, ..., n_m$ responses of categories 0, 1, ..., m to the persons in ascending (x_G) or descending (x_A) order (see Table 1).

An item can be evaluated according to its fit to any trait or dimension without knowing or estimating its parameters (e.g., when new items are answered by individuals with known β levels). If the β levels are not known beforehand, the item parameters have to be estimated first in order to estimate the βs. In any case, the item parameters are not directly involved in the index.

Q_i also can be seen as a probabilistic item discrimination index that makes no use of Pearson's correlation coefficient but uses only response frequencies. $Q_i = 0$ indicates perfect item discrimination (high β levels with high-scoring responses), whereas $Q_i = 1$ indicates a perfect negative item discrimination (high β levels with low-scoring responses). $Q_i = .5$ indicates independence of the trait and the item.

Table 1
Observed Responses (x_{obs}), Guttman Responses (x_G), and Anti-Guttman Responses (x_A) for 18 Persons, 3 Response Categories ($x = 0, 1, 2$), and Item Scores $f(0) = 4, f(1) = 8, f(2) = 6$

β_v	x_{obs}	x_G	x_A
−3.2	0	0	2
−2.8	0	0	2
−2.8	1	0	2
−1.2	0	0	2
−1.2	0	1	2
−1.2	1	1	2
.3	1	1	1
.3	1	1	1
.3	1	1	1
.3	2	1	1
1.2	1	1	1
1.2	1	1	1
1.2	1	2	1
1.2	2	2	1
2.0	2	2	0
2.0	2	2	0
2.0	2	2	0
3.7	2	2	0

Performance of Q_i

To investigate the performance of Q_i, simulation studies were implemented in which the fit of a single item to the Rasch model or partial credit model was successively deteriorated. The simulation studies first specified an item pattern that had maximum fit to the model (i.e., with a perfect Guttman pattern). Then a certain proportion of (randomly selected) responses (1/10, 2/10, ..., 5/10, 10/10) were replaced with random responses that were independent of the person's β level. In each step the density of Q_i was calculated.

Figure 1 shows the results of six different sets of model parameters. The β distributions in Figures 1a–1e were normal. In these cases, the standard deviation of β was arbitrary because it could be cancelled out when computing Q_i (see Equation 15, where a constant factor of β can be reduced). In contrast to the fit statistic $Z(Q_i)$ (see below) in which the power strongly depends on the variance of β, a stretching or squeezing of the β distribution by means of a constant factor is irrelevant for the descriptive measure Q_i. In Figure 1f, the

β distribution was not normal but bimodal in order to investigate the effects of extreme distributions on Q_i.

The β level of persons (N = 200 and 1,000) and the difficulty of the items shown in Figure 1 were jointly defined by the category probabilities $p_x = n_{ix}/N$. Figures 1a and 1b show the results from dichotomous items; Figure 1c–1f show those from items with four categories for which the partial credit model was applied.

Figures 1a–1e show that the expectation of Q_i under chance conditions (10/10) is $Q_i = .5$; the function located to the far right corresponds to this chance condition. The other functions in these figures correspond to different degrees of disturbance (i.e., 1/10 to 5/10). All functions were parallel—a higher degree of disturbance or misfit produced a higher expectation of Q_i and all Q_i distributions had approximately the same variance for a given sample size. Furthermore, the distances between the five functions corresponding to the first five proportions of random individuals (1/10 to 5/10) decreased slightly but consistently.

There are only small differences among Figures 1a–1f. Figures 1a and 1b show the performance of Q_i for two dichotomous items with different difficulty levels (relative solution frequencies were $p = .6$ and $p = .1$, respectively). The functions in Figure 1a are steeper than the functions in Figure 1b. The variance of Q_i is smaller if the item difficulty fits the mean β level of the individuals (Figure 1a with a mean solution probability $p = .6$). In the case of extreme items (Figure 1b) the variances are somewhat higher so the functions are not as steep. However, the expectations of Q_i were identical for items with different difficulties.

The same result also can be seen in Figures 1c and 1d. In these figures, the items had four categories and the partial credit model was used. Figure 1d corresponds to a more extreme item and has slightly higher variances of Q_i but constant expectations.

Figures 1e and 1f show a variation in the sample size (Figure 1e) and in the range of β levels (Figure 1f). The only effect of increasing the sample size was on the variances of Q_i, which became smaller. The broad β range (Figure 1f) caused the variance of Q to become slightly smaller than that of Figure 1c.

These data show that Q_i, under different degrees of model violation, had an expectation that reflected the degree of violation and that was not affected by the item difficulty, sample size, or β range. Q_i also had a variance that was constant for different degrees of violation and that became smaller for larger sample sizes and items with difficulty that matched β. Q_i allows for a comparative evaluation of all items in a test. Each person contributes to this measure of item fit by his/her deviation from his/her optimum response, but without calculating score residuals. Statistical inference about the model fit of single items, however, requires the derivation of a statistic with a known (asymptotical) distribution under model assumptions.

Statistical Inference With Q_i

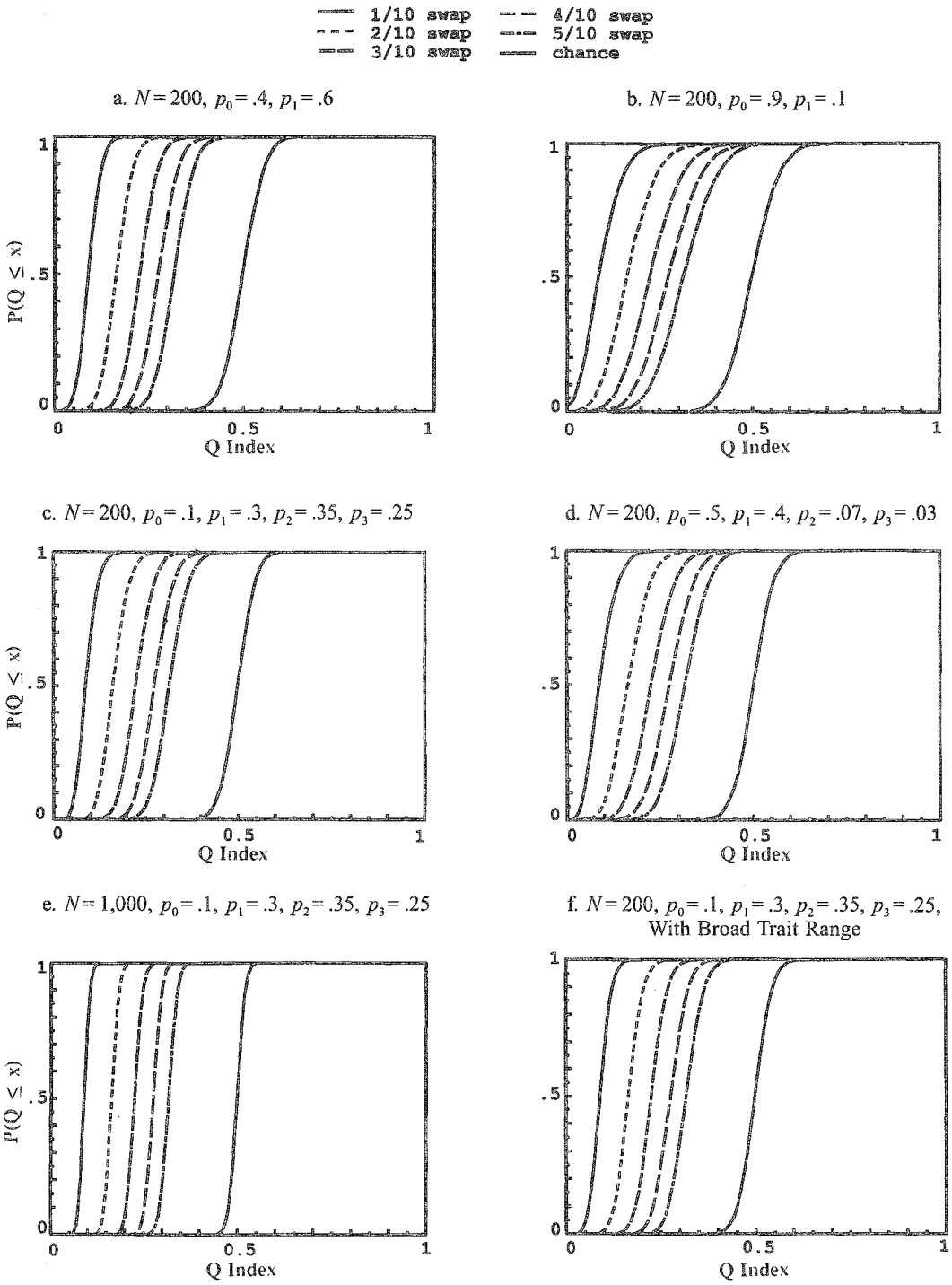
The previous section demonstrated that Q_i is symmetrically distributed in the (0,1) interval. The .5 midpoint indicates random response behavior. However, this does not provide information about the distribution of Q_i under the assumptions of the Rasch model or the partial credit model. In fact, under model assumptions an asymptotic normal distribution can be derived for the numerator of Q_i. (The denominator does not depend on the observed data and, hence, does not contribute to the variance of Q under model assumptions.)

The denominator of Q_i is the log-likelihood ratio of the Guttman and anti-Guttman patterns (Equation 14) and, hence, is a normalization factor that is independent of the observed pattern. All variation in Q_i caused by the data is introduced in the numerator, which is the log-likelihood ratio of the observed and the Guttman pattern. This log-likelihood ratio,

$$Q_1^* = \log \mathrm{LR}_{i,G} = \sum_{v=1}^{N} (x_{vi} - x_{v,G})\beta_v \ , \qquad (16)$$

is a weighted sum of all person parameters, β. It is known from maximum likelihood theory that the estimate, $\hat{\beta}$, for each person is asymptotically normally distributed. Assuming that all individuals respond independently, the weighted sum (Equation 16) is also asymptotically normally distributed with expectation

Figure 1
Cumulated Probability Distributions of Q_i Under Six Degrees of Disturbance Starting From a Perfect Pattern



$$\mathrm{E}(Q_i^*) = \sum_{v=1}^{N}\left[\mathrm{E}(x_v\beta_v) - x_{v,\mathrm{G}}\beta_v\right] = \sum_{v=1}^{N}\left[\sum_{x=0}^{m} p_{vi}(x)x\beta_v\right] - x_{v,\mathrm{G}}\beta_v \tag{17}$$

and variance

$$\mathrm{V}(Q_i^*) = \sum_{v=1}^{N}\mathrm{V}(x_v\beta_v - x_{v,\mathrm{G}}\beta_v) = \sum_{v=1}^{N}\mathrm{V}(x_{v,\mathrm{G}}\beta_v)\,. \tag{18}$$

The variance of $x_v\hat{\beta}_v$ is the variance of the product of two random variables, because $\hat{\beta}_v$ is only an estimate of the true trait level. This random variable has the expectation $\beta_v = \mathrm{E}(\hat{\beta}_v)$ and is asymptotically normally distributed when maximum likelihood estimators are applied.

In order to not underestimate the variance of $x_v\hat{\beta}_v$ due to measurement error, Equation 19 can be used. The expectation β_v is estimated by $\hat{\beta}_v$, and its variance is estimated by the squared standard error of measurement, $S_e^2(\hat{\beta}_v)$, which is estimated using the information function. If it is assumed that x_{vi} is independent of the error of measurement of $\hat{\beta}_v$ (e.g., when β is estimated without item i or the number of items is sufficiently large), it follows that (Kendall & Stuart, 1973, p. 245),

$$\begin{aligned}
\mathrm{V}(x_{vi}\hat{\beta}_v) &= \mathrm{E}(\hat{\beta}_v^2 x_{vi}^2) - \mathrm{E}(\hat{\beta}_v x_{vi})^2 \\
&= \mathrm{E}(\hat{\beta}_v^2)\mathrm{E}(x_{vi}^2) - \mathrm{E}(\hat{\beta}_v)^2\mathrm{E}(x_{vi})^2 \\
&= \left[\mathrm{E}(\hat{\beta}_v^2) - \mathrm{E}(\hat{\beta}_v)^2 + \mathrm{E}(\hat{\beta}_v)^2\right]\left[\mathrm{E}(x_{vi}^2 - \mathrm{E}(x_{vi})^2 + \mathrm{E}(x_{vi})^2\right] - \mathrm{E}(\hat{\beta}_v)^2\mathrm{E}(x_{vi})^2 \\
&= \left[\mathrm{V}(\hat{\beta}_v) + \mathrm{E}(\hat{\beta}_v)^2\right]\left[\mathrm{V}(x_{vi}) + \mathrm{E}(x_{vi})^2\right] - \mathrm{E}(\hat{\beta}_v)^2\mathrm{E}(x_{vi})^2 \\
&= \mathrm{V}(\hat{\beta}_v)\mathrm{V}(x_{vi}) + \mathrm{E}(\hat{\beta}_v)^2\mathrm{V}(x_{vi}) + \mathrm{V}(\hat{\beta}_v)\mathrm{E}(x_{vi})^2\,,
\end{aligned} \tag{19}$$

where

$$\mathrm{E}(x_{vi}) = \sum_{x=0}^{m} p_{vi}(x)x \tag{20}$$

and

$$\mathrm{V}(x_{vi}) = \sum_{x=0}^{m} p_{vi}(x)[x - \mathrm{E}(x)]^2\,. \tag{21}$$

The probability, $p_{vi}(x)$, of a single response x by individual v on item i can only be calculated by means of person and item parameter estimates using Equation 8.

Standardizing Q_i^* by means of its expectation and variance

$$Z(Q_i) = \frac{Q_i^* - \mathrm{E}(Q_i^*)}{\mathrm{V}(Q_i^*)^{1/2}}\,, \tag{22}$$

a Z statistic is obtained that can be used to decide whether an item pattern deviates significantly from the model. Note that in all calculations the uncorrected β level parameter estimates must be used. Applying the correction term $(k-1)/k$, where k is the number of items, often used for unconditional maximum likelihood estimators (e.g., Wright & Douglas, 1977) causes a systematic bias of the $Z(Q_i)$ statistic. The reason for this is that only the uncorrected estimates satisfy the likelihood function: the $\hat{\beta}$s are unconditional estimators, even if the item parameters have previously been estimated conditionally.

Furthermore, this derivation is based on the assumption of independence of x_{vi} and $\hat{\beta}_v$ (i.e., the error of measurement of the latter). In the case of a covariance of both variables, the variance of Q_i^* is underestimated and, hence, the test of significance becomes stronger so the null hypothesis is rejected earlier.

In general, the power of this item-fit statistic depends on the range of β levels in the sample, as do other item-fit measures based on item response residuals. As Andrich (1988) has shown, power decreases when persons are well-targeted to an item and there is little spread of the β levels. In this case both responses of a dichotomous item are equally likely, so that misfit cannot be revealed by residual analysis or by testing the pattern likelihood.

Example Application

This example shows the sensitivity of Q_i for both item and person misfit. A questionnaire was analyzed that was part of the 1985 Shell study (Jugendwerk der Deutschen Shell, 1985). The 10-item questionnaire measures the construct "adolescent centrism" that reflects the attitude of young people toward the older generation. The translated items are listed in Table 2.

Each item was rated on a four-point scale from *don't agree* (scored 0) to *strongly agree* (scored 3) by $N = 1{,}472$ individuals. Using the mixed Rasch model (Rost, 1990, 1991), 76% of the individuals responded in a manner that was compatible with the Rasch model ("scalables") and approximately 24% of the sample responded in a way that was incompatible with the assumptions of the ordinal Rasch model ("unscalables;" Rost & Georg, 1991). This latter phenomenon was attributed to the conditions of field research, in which the problem of aberrant response patterns is more serious than under laboratory or school conditions.

The mean Q_i values were smaller for the "scalables" than for the "unscalables" (see Table 3). According to the $Z(Q_i)$ values for the 10-item solution, Item 6 fit poorly and should be removed from the attitude scale. In fact, this item violated a basic rule of item construction, which is that only one assertion should be addressed at a time (trying to understand parents is one thing, evaluating it as difficult is another).

For the 9-item analysis (eliminating Item 6), despite the fact that for the unscalables none of the items deviated significantly from the model, all Q_i values were relatively large. This demonstrates the use of a descriptive measure other than the significance test for single items. Whereas the former is sensitive to the general disturbance in the unscalables, the latter evaluates the deviation of single items under the null hypothesis that the model holds for all items. In such a group of unscalables, however, there are no single deviating items, but a general high level of noise.

This example is atypical for illustrating a measure of item fit, because the improvement of model fit primarily is achieved by selecting "aberrant" persons, not items. On the other hand, the improvement of

Table 2
Translation of the 10 Items on the Adolescent Centrism Scale

1.	The police force treats adolescents unfairly.
2.	In our society you are confronted with animosity everywhere—which is completely demoralizing.
3.	Our society really does do a lot for young people.
4.	Young people put their foot down when necessary and do not put up with everything at work.
5.	I really owe a lot to my parents.
6.	I try to understand my parents—even when it is difficult.
7.	Very few adults really understand young people's problems.
8.	I don't believe in adults' experience—I prefer to depend on myself.
9.	I learn more from friends my own age than from my parents.
10.	Parents always interfere in things that are none of their business.

Table 3
Q_i Values for 10 Items of Adolescent Centrism in the Total Group, the "Scalable" and "Unscalable" Subgroups, and for the Reduced 9-Item Questionnaire

	10 Items						9 Items					
	Total Group		76% Scalables		24% Unscalables		Total Group		75% Scalables		25% Unscalables	
Item	Q_i	$Z(Q_i)$	Q_i	$Z(Q_i)$	Q_i	$Z(Q_i)$	Q_i	$Z(Q_i)$	Q_i	$Z(Q_i)$	Q_i	$Z(Q_i)$
1	.20	−.5	.18	−.3	.28	.8	.20	−.1	.17	.2	.26	.4
2	.20	−1.0	.18	−1.0	.25	.1	.20	−.8	.17	−.8	.24	−.3
3	.22	−.1	.20	.1	.24	0.0	.22	.2	.19	.3	.24	−.3
4	.22	1.2	.20	1.3	.26	.3	.21	1.3	.19	1.3	.26	.5
5	.19	−1.5	.18	−1.6	.18	−1.1	.21	.8	.19	.8	.24	0.0
6	.29	3.6	.28	3.2	.30	1.3	-	-	-	-	-	-
7	.24	1.0	.21	.5	.28	.6	.22	.7	.19	.3	.25	.5
8	.18	−1.0	.16	−.9	.20	−.8	.18	−.7	.16	−.8	.22	−.3
9	.19	−.7	.17	−.4	.19	−1.4	.18	−.8	.16	−.6	.21	−.8
10	.19	−.7	.17	−.5	.25	.2	.19	−.7	.16	−.7	.24	.2

model fit using the selection of items with a high Q_i is evident and needs no demonstration. This example shows that high Q_i values may reflect too much noise in the data or, in this case, a certain number of misfitting persons.

Discussion

Test construction requires information about single items; for example, their difficulties, their threshold distances, and their relation to the latent trait. The latter usually is provided by an item discrimination index such as the item-total correlation in classical test theory or the slope of the IRF in the two-parameter and three-parameter logistic models. The Rasch model has none of these, because IRFs are assumed to be parallel. Q_i provides information about item discrimination when dichotomous and polytomous Rasch models are used. Q_i is low when high β level persons have higher item scores and when low β level persons have lower scores. Q_i is computed without the use of correlations and without score residuals. Instead, Q_i is a probabilistic measure of item fit for dichotomous and ordinal Rasch models based on the likelihood of the item patterns. It provides both a descriptive measure of the degree of disturbance in the data and a measure for statistical inference about the fit of single items.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123–140.

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

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47,* 105–113.

Andrich, D. (1988). *Rasch models for measurement.* London: Sage.

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

Drasgow, F., Levine, M., & Williams, E. (1985). Appropriateness measurement with polychotomous item response models and standardized residuals. *British Journal of Mathematical and Statistical Psychology, 38,* 67–86.

Glas, C. A. W. (1989). *Contributions to estimating and testing Rasch models.* Unpublished doctoral dissertation, The Hague, CIP-Gegevens Koninglijke Bibliothek, University of Twente, The Netherlands.

Jugendwerk der Deutschen Shell (Ed.). (1985). *Jugendliche und Erwachsene '85. Generationen im Vergleich* [Adolescents and adults 1985. A comparison of generations]. Opladen: Author.

Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistics* (Vol. 1). London: Griffin.

Levine, M. V., & Drasgow, F. (1982). Appropriateness

measurement: Review, critique, and validating studies. *British Journal of Mathematical and Statistical Psychology, 35,* 42–56.

Levine, M. V., & Rubin, D. F. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269–290.

Martin-Löf, P. (1973). Statistical models. Notes from Seminars, 1969–70. (by Rolf Sundberg)

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 147–174.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9,* 49–57.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 48,* 49–72.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14,* 127–137.

Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. *Applied Psychological Measurement, 12,* 397–409.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271–282.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology, 44,* 75–92.

Rost, J., & Georg, W. (1991). Alternative Skalierungsmöglichkeiten zur klassischen Testtheorie am Beispiel der Skala "Jugendzentrismus" [Scaling models as an alternative to classical test theory, exemplified with the scale "adolescent centrism"]. *Zentralarchir Information, 28,* 52–75.

Rost, J., & von Davier, M. (1992). *MIRA: A PC-program for the mixed Rasch model* [Computer program]. Kiel, Germany: Institute for Science Education–IPN.

Tarnai, C., & Rost, J. (1990). *Identifying aberrant response patterns in the Rasch model: The Q index.* Münster, Germany: Sozialwissenschaftliche Forschungsdokumentationen.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7,* 81–96.

van den Wollenberg, A. L. (1979). *The Rasch model and time limit tests.* Nijmegen, The Netherlands: Studentenpers.

van den Wollenberg, A. L. (1982). A simple and effective method to test the dimensionality axiom of the Rasch model. *Applied Psychological Measurement, 6,* 83–92.

van den Wollenberg, A. L. (1988). Testing a latent trait model. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 31–50). New York: Plenum.

Wright, B. D. (1980). Afterword. In G. Rasch, *Probabilistic models for some intelligence and attainment tests* (pp. 185–199). Chicago: University of Chicago Press.

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1,* 281–294.

Wright, B. D., & Masters, G. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23–48.

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

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245–262.

Author's Address

Send requests for reprints or further information to Jürgen Rost or Matthias von Davier, Institute for Science Education–IPN, Olshausenstrasse 62, D-24098 Kiel, Germany. Internet: rost@ipn.uni-kiel.de or vdavier@ipn.uni-kiel.de.