Journal of Educational Measurement
Winter 2006, Vol. 43, No. 4, pp. 335–353

Modeling Randomness in Judging Rating Scales with a Random-Effects Rating Scale Model

Wen-Chung Wang
National Chung Cheng University, Chia-Yi, Taiwan
Mark Wilson
University of California, Berkeley
Ching-Lin Shih
National Chung Cheng University, Chia-Yi, Taiwan,
National Taichung University, Taichung, Taiwan

This study presents the random-effects rating scale model (RE-RSM) which takes into account randomness in the thresholds over persons by treating them as random-effects and adding a random variable for each threshold in the rating scale model (RSM) (Andrich, 1978). The RE-RSM turns out to be a special case of the multidimensional random coefficients multinomial logit model (MRCMLM) (Adams, Wilson, & Wang, 1997) so that the estimation procedures for the MRCMLM can be directly applied. The results of the simulation indicated that when the data were generated from the RSM, using the RSM and the RE-RSM to fit the data made little difference: both resulting in accurate parameter recovery. When the data were generated from the RE-RSM, using the RE-RSM to fit the data resulted in unbiased estimates, whereas using the RSM resulted in biased estimates, large fit statistics for the thresholds, and inflated test reliability. An empirical example of 10 items with four-point rating scales was illustrated in which four models were compared: the RSM, the RE-RSM, the partial credit model (Masters, 1982), and the constrained random-effects partial credit model. In this real data set, the need for a random-effects formulation becomes clear.

Rating scales are those that require respondents to indicate their response choices from an ordered series of categories. For instance, they might rate the frequency with which each of the behaviors has occurred within a month (e.g., never, a few times, often, and most of the time). Within the framework of item response theory (IRT), the rating scale model (RSM) (Andrich, 1978), which is an extension of the Rasch model (Rasch, 1960), is commonly used to fit the responses to rating scale items (Andrich, 1988; Bond & Fox, 2001; Embretson & Reise, 2000; Smith & Smith, 2004; Wright & Masters, 1982). For dichotomous items, under the Rasch model the probability of a positive response (or a correct answer; scoring 1) to item i for a person with latent trait θ_n is

$$p_{ni1} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}, \tag{1}$$

where δ_i is the difficulty of item i. Equation (1) can be expressed as

$$\log(p_{ni1}/p_{ni0}) = \theta_n - \delta_i, \tag{2}$$

where p_{ni0} is the probability of scoring 0 (a negative response or an incorrect answer).

When item responses are scored polytomously according to a common rubric, such as rating scale items in which all items have the same rating scale structure, Equation (2) can be extended to

$$\begin{aligned}\log(p_{nij}/p_{ni(j-1)}) &= \theta_n - \delta_{ij}\\ &\equiv \theta_n - (\delta_i + \tau_j),\end{aligned} \tag{3}$$

where p_{nij} and $p_{ni(j-1)}$ are the probabilities of scoring j and $j-1$ on item i for person n, respectively; δ_{ij} is the j-th step difficulty of item i; after reparameterization δ_i is the difficulty (location parameter) of item i; and τ_j is the j-th threshold parameter (also called the intersection parameter) for every item. Equation (3) is known as the RSM, in which all items are constrained to share the same set of threshold parameters in order to embody the common scoring rubric used throughout the test.

Polytomous items are not always scored according to a common rubric. For example, each constructed-response item in an ability test usually has its own scoring rubric. Hence, it is reasonable to release the constraint in the RSM and allow each item to have its own set of threshold parameters so that Equation (3) is extended to

$$\begin{aligned}\log(p_{nij}/p_{ni(j-1)}) &= \theta_n - \delta_{ij}\\ &\equiv \theta_n - (\delta_i + \tau_{ij}),\end{aligned} \tag{4}$$

where τ_{ij} is the j-th threshold parameter for item i, and the others are defined as in the RSM. Equation (4) is called the partial credit model (PCM) (Masters, 1982).

The RSM originates in attitude surveys where the respondent is presented the same response choices for a set of items so that all items have the same rating scale structure. The motivation for the PCM derives from ability tests where responses that are incorrect, but indicate some knowledge, are given partial credit towards a correct response. The amount of partial correctness often varies over items so that each item has its own rating scale structure (Wright, 1999). Although the idea of the RSM rather than the PCM matches better the motivation for rating scale items (Wright & Masters, 1982), it is quite possible that respondents are not treating the categories in the same way across all the items—the PCM provides an alternative that examines this possibility. Note that τ_{ij} in the PCM is a fixed effect, meaning that the threshold parameter remains constant over persons. Employing the PCM to analyze rating scale items for better fit is referred to as the "fixed-effects" approach, because additional fixed-effects parameters are added to the RSM to form the PCM.

Rating scale items usually require respondents' subjective judgments, which are very likely to vary over persons. In such a case, the RSM and the PCM may be too stringent to fit real data from rating scale items, because the subjective nature in judging rating scale items is not taken into account appropriately. Acknowledging the subjective nature of the judgment, practitioners usually adopt more liberal criteria in assessing model-data fit when fitting the RSM to rating scales than when fitting the Rasch model or the PCM to ability tests. For instance, Wright, Linacre, Gustafson, and Martin-Löf (1994) suggested using a wider critical range of (0.6, 1.4) on the mean square item fit statistics for rating scale items, but a narrower range of (0.8, 1.2) for high-stakes ability items.

In our approach, we instead make the model more complex in order to account for the subjective nature in judging rating scale items. That is, we acknowledge the subjective nature by directly taking it into account in the model. We treat the threshold parameter as a random effect over persons by imposing a distribution onto it and thus form the so called random-effects rating scale model (RE-RSM), in which

$$\log(p_{nij}/p_{ni(j-1)}) = \theta_n - (\delta_i + \tau_{nj}) \tag{5}$$

and

$$\tau_{nj} \sim N\left(\tau_j, \sigma_j^2\right). \tag{6}$$

Equations (5) and (6) can be reformulated as

$$\log(p_{nij}/p_{ni(j-1)}) = \theta_n - (\delta_i + \tau_j) + \gamma_{nj} \tag{7}$$

and

$$\gamma_{nj} \sim N\left(0, \sigma_j^2\right). \tag{8}$$

The variance σ_j^2 depicts the amount of the variations in subjective judgment on the j-th threshold over persons. The larger the variance, the more variation in the subjective judgment, and thus the less consistent the rating scale over the persons. If

$$\sigma_j^2 = 0, \quad \text{for all } j, \tag{9}$$

then the RE-RSM reduces to the RSM. Because the random threshold variables (γs) are used to describe the random nature of subjective judgment, they are assumed to be independent of each other and of the latent trait θ. For identification of the parameters, the means of the latent trait and the random threshold variables are set at zero.

A similar extension can be applied to the PCM so as to form the random-effects partial credit model (RE-PCM):

$$\log(p_{nij}/p_{ni(j-1)}) = \theta_n - (\delta_i + \tau_{ij}) + \gamma_{nij} \tag{10}$$

and

$$\gamma_{nij} \sim N(0, \sigma_{ij}^2). \tag{11}$$

Assume there are I items in a test, each with J response categories. Under the RE-RSM and the RE-PCM, there will be $J - 1$ and $I \times (J - 1)$ variances for the random thresholds, respectively. Under most regular testing conditions, the RE-PCM is not identifiable. Only if data are sufficient, such as longitudinal data in which the test is repeatedly administered, would the parameters in the RE-PCM be identifiable and calibrated accurately. In practice, some constraint has to be made to reduce the number of parameters in the random thresholds, such as constraining the random thresholds to have equal variances over items:

$$\gamma_{nij} \sim N\left(0, \sigma_j^2\right). \tag{12}$$

This is called the constrained random-effects partial credit model (CRE-PCM), in which only $J - 1$ variances for the random thresholds are estimated.

Figure 1 represents graphically the RSM, the PCM, the RE-RSM, the RE-PCM, and the CRE-PCM for a test with 10 items each having five points. Under the RSM and the PCM, δ_{ij}, presented as a point, is treated as a fixed effect. The lines for the

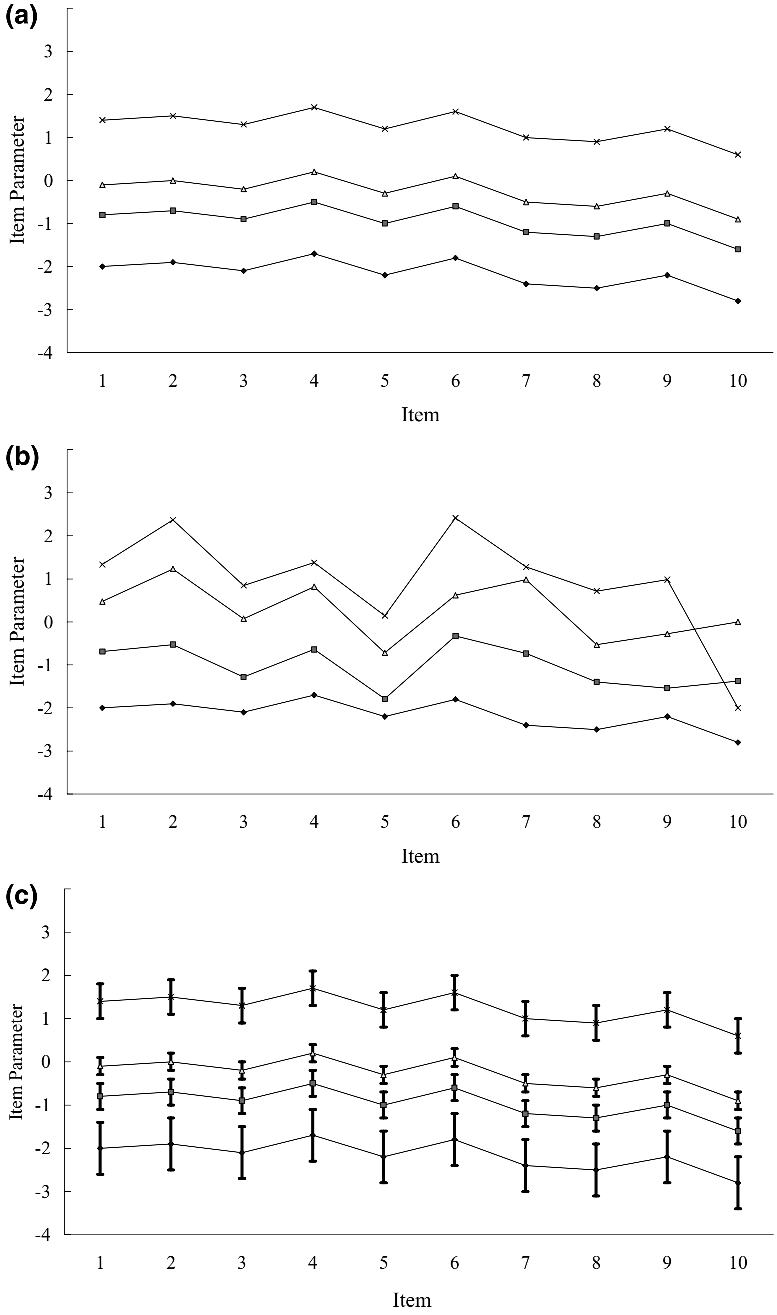


FIGURE 1. *Graphic representations for (a) the rating scale model, (b) the partial credit model, (c) the random-effects rating scale model, (d) the random-effects partial credit model, and (e) the constrained random-effects partial credit model, for 10 items each with five points.*

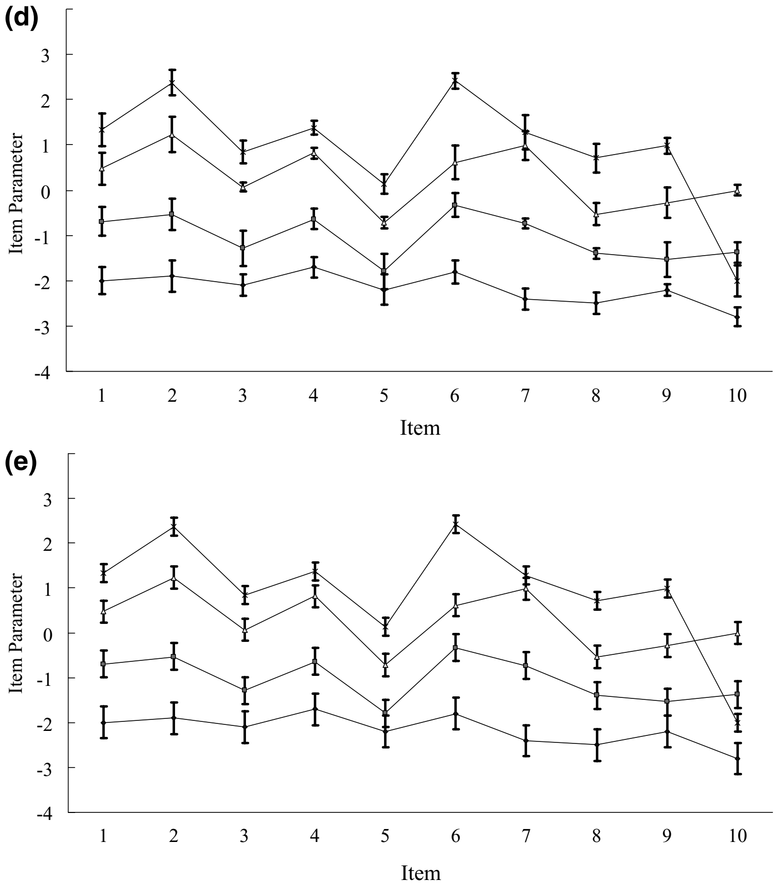


FIGURE 1. *Continued.*

four thresholds are constrained to be parallel under the RSM, indicating all the items share the same set of threshold parameters, whereas no parallel constraint is made under the PCM, meaning that each item has its own set of threshold parameters. Under the other three random effects models, δ_{ij}, presented as a point plus a band, is treated as a random effect. The widths of the bands indicate the magnitudes of variations in the thresholds. Under the RE-RSM and the CRE-PCM, the widths of the bands are constrained to be equal over items, where those under the RE-PCM are not. The larger the widths, the more variations in the thresholds will be across persons. As a result, some categories are endorsed substantially less than the other so that they provide little information about persons. Hence, the test reliability will be decreased.

Treating item parameters as random effects (also known as hierarchical or multilevel modeling; Adams, Wilson, & Wu, 1997; Fox & Glas, 2001; Mislevy & Bock, 1989; Patz & Junker, 1999) has been adopted in several testing situations. For example, when items are connected by common stimuli (e.g., a reading comprehension passage or a figure, called a testlet or an item bundle; Rosenbaum, 1988; Wainer & Kiely, 1987), fitting standard item response models to testlet-based responses ignores the possible local item dependence between items within a testlet, which tends to overestimate the precision of person measures (i.e., test reliability), and yields biased estimation for item difficulty and discrimination parameters (Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Wainer & Lukhele, 1997; Wainer & Thissen, 1996; Wainer & Wang, 2000; Wilson & Adams, 1995; Yen, 1993). To resolve this problem, several researchers (e.g., Bradlow, Wainer, & Wang, 1999; Rijmen & De Boeck, 2002; Wainer, Bradlow, & Du, 2000; Wang, Bradlow, & Wainer, 2002; Wang & Wilson, 2005a, 2005b) extended standard item response models by including additional random variables for local item dependence within testlets to capture the interaction between persons and items within testlets. The present study is similar to these studies in that the random-effects approach to item parameters is used to describe randomness in item parameters, but is different from them in that the present study focuses on randomness within thresholds, whereas the previous studies focus on randomness within testlets.

Parameter Estimation

When the RE-RSM or the CRE-PCM is fitted to rating scale items, there are altogether one latent trait and J – 1 random threshold variables (J is the number of points in rating scale items). Assuming

$$\theta_n \sim N\left(0, \sigma_\theta^2\right), \tag{13}$$

one can define $\boldsymbol{\theta}_n^{\mathrm{T}} = \left[\theta_n, \gamma_{n1}, \ldots, \gamma_{n,J-1}\right]$, which follows a multivariate normal distribution $N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a diagonal matrix because all the random variables are constrained to be independent. Viewing the θ and γs this way, one finds that the RE-RSM and the CRE-PCM (as well as the RE-PCM) are both multidimensional item response models in which the multiple dimensions are constrained to be independent. Moreover, it can be shown that the RE-RSM and the CRE-PCM are both special cases of the multidimensional random coefficients multinomial logit model (MRCMLM; Adams et al., 1997) so that the parameters in these models can be estimated using the computer program ConQuest (Wu, Adams, & Wilson, 1998). No extra efforts are needed to derive parameter estimation procedures or to develop computer software specifically for the RE-RSM or the CRE-PCM.

Let person n's levels on the L latent traits be denoted as $\boldsymbol{\theta}_n^{\mathrm{T}} = (\theta_{n1}, \ldots, \theta_{nL})$, which are considered to represent a random sample from a population with a multivariate density function $g(\boldsymbol{\theta}_n; \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ indicates a vector of parameters that characterize the distribution. In this study, g is constrained to be multivariate normal so that $\boldsymbol{\alpha} \equiv (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Under the MRCMLM, the probability of a response in category j of item i for a person with latent traits $\boldsymbol{\theta}_n$ is

$$p_{nij} = \frac{\exp\left(\mathbf{a}_{ij}^{\mathrm{T}}\boldsymbol{\theta}_n + \mathbf{b}_{ij}^{\mathrm{T}}\boldsymbol{\xi}\right)}{\sum_{k=1}^{J_i} \exp\left(\mathbf{a}_{ik}^{\mathrm{T}}\boldsymbol{\theta}_n + \mathbf{b}_{ik}^{\mathrm{T}}\boldsymbol{\xi}\right)}, \tag{14}$$

where J_i is the number of categories in item i; $\boldsymbol{\xi}$ is a vector of difficulty parameters that describe the items; $\mathbf{a}_{ij}$ is a score vector given to category j of item i across the L latent traits; and $\mathbf{b}_{ij}$ is a design vector given to category j of item i that describes the linear relationship among the elements of $\boldsymbol{\xi}$.

Using $\mathbf{a}_{ij}$ and $\mathbf{b}_{ij}$ to specify the relationship between items and persons allows a general model to be written that includes most of the existing unidimensional Rasch models and many multidimensional models, such as multidimensional forms of the Rasch model, the RSM, the PCM, the RE-RSM, and the CRE-PCM as described in this article. A necessary and sufficient condition for identification of the parameters in the MRCMLM has been derived (Volodin & Adams, 1995; Wu et al., 1998). ConQuest is implemented with marginal maximum likelihood estimation and Bock and Aitkin's (1981) formulation of the EM algorithm. In addition to the standard technique, which uses fixed quadrature points for integration, ConQuest also provides a Monte Carlo method where the quadrature points are readjusted according to recent estimates. This Monte Carlo approach is more feasible than the fixed quadrature approach when the number of dimensions is large (e.g., larger than three). After the model parameters $\boldsymbol{\mu}$, Σ and $\boldsymbol{\xi}$ are calibrated, point estimates for individual persons can be obtained from the mean vector of the marginal posterior distribution, called the expected a posteriori (EAP) estimates (Bock & Mislevy, 1982), or the maximum point of the conditional likelihood, called the maximum likelihood estimates.

The MRCMLM, being a member of the exponential family of distributions, can be viewed as a generalized linear mixed model (De Boeck & Wilson, 2004; McCullagh & Nelder, 1989; McCulloch & Searle, 2001; Nelder & Wedderburn, 1972; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). In addition to ConQuest, the SAS NLMIXED procedure (SAS Institute, 1999; Wolfinger & SAS Institute, n.d.) is an alternative for fitting many common nonlinear and generalized linear mixed models, including the MRCMLM. According to the authors' experiences with multidimensional and random-effects modeling, ConQuest takes only a few minutes to converge, whereas NLMIXED may take several hours to converge (or sometimes even fail to converge). Thus, ConQuest was used for all analyses in this study.

As ConQuest had never been applied to the RE-RSM, parameter recovery and effects of model misspecification between the RSM and the RE-RSM were unknown. Simulations were thus conducted to investigate how well the generating parameters could be recovered, when the generating model and the analysis model were identical (i.e., both were the RE-RSM or the RSM) or different (i.e., the generating model was the RE-RSM but the analysis model was the RSM, or the other way around). Below, the design, data analysis and results of the simulations are described. An empirical example of four-point rating scales is also given in which four models were compared: the RSM, the PCM, the RE-RSM, and the CRE-PCM.

Method

Design

Two conditions were investigated through simulation. Under the first condition, the RSM was used to generate item responses of 15 five-point rating scale items by 500 persons. The difficulties for the 15 items ranged uniformly from –1.5 to 1.5. The four threshold parameters were –2.0, –.5, .5, and 2.0, respectively. The person parameters were generated from $N(0, 1)$. These generating values were arbitrarily chosen, but they are within the usual scope of values we have observed. The generated data were analyzed using the RSM and the RE-RSM, respectively. It was hypothesized that (a) using the RSM would recover the generating parameters very well, (b) using the RE-RSM would not result in statistically better fit to the data than using the RSM, and (c) the variance estimates for the random thresholds under the RE-RSM would be very close to zero.

The second condition was identical to the first one, except that the generating model was the RE-RSM in which the four random thresholds had variances of .75, 2.00, 1.50, and 3.00, respectively, which represented median to large random-effects. The generated data were analyzed using the RSM and the RE-RSM, respectively. It was hypothesized that (a) using the RE-RSM could recover the generating parameters very well, and (b) using the RSM would result in biased estimates and poor model-data fit because the randomness in the thresholds were ignored. One hundred replications were made under each condition.

Data Generation

A FORTRAN 90 computer program was written by the authors to generate item responses. The generating procedure contained the following steps: (a) The latent trait and random threshold variables (θ and γ) were randomly generated independently from a normal distribution with mean zero; (b) these generated latent trait and random threshold variables and the defined item parameters were used to compute the corresponding category probability and the cumulative probabilities using Equation (7); (c) these cumulative probability values were compared to a random number from a uniform (0, 1) distribution. The simulated item response was defined as the highest score category at which the random number was less than or equal to the associated cumulative probability.

Analysis

The generated data sets were calibrated using ConQuest. The Monte Carlo approach with 2000 nodes was used for integration. The other options were set at the default (e.g., maximum numbers of the Newton iterations in the M-step and the E-step were 10 and 1,000, respectively; iterations terminated when maximum parameter or deviance change was less than .0001). The bias and sampling variance for the estimator $\hat{\varsigma}$ were computed as

$$\text{Bias}(\hat{\varsigma}) = \sum_{r=1}^{100} (\hat{\varsigma}_r - \varsigma)/100, \tag{15}$$

$$\text{Var}(\hat{\varsigma}) = \sum_{r=1}^{100} (\hat{\varsigma}_r - \bar{\hat{\varsigma}})^2/99, \tag{16}$$

respectively, where ς and $\bar{\hat{\varsigma}}$ were the generating value and the mean estimate across replications, respectively. The Z statistic for testing whether the estimator was biased was computed as

$$Z(\hat{\varsigma}) = \text{Bias}(\hat{\varsigma})/\sqrt{\text{Var}(\hat{\varsigma})/100}. \quad (17)$$

If the computed Z was beyond the .01 significance level (±2.576), the corresponding estimator was declared as biased.

In addition to these statistics, the weighted and unweighted item fit statistics (also called infit and outfit item statistics, respectively; Wright & Masters, 1982; Wu et al., 1998) were computed to compare model-data fit of the RSM and the RE-RSM. In theory, when the data fit the model's expectation, the infit and outfit mean square errors (MSE) will have a mean of unity, and their corresponding t statistics will approximately follow the standard normal distribution. The likelihood deviances ($-2 \times$ loglikelihood) for the two models, G^2_{RSM} and $G^2_{\text{RE-RSM}}$, were computed to compare global model-data fit between models. Because the RE-RSM has four more parameters than the RSM, the difference in the two likelihood deviances was tested against the χ^2 distribution with four degrees of freedom, which is conservative (Stram & Lee, 1994, 1995).

Results

Condition 1: No Random Threshold Effects

The generating model was the RSM and the data were analyzed using the RSM and the RE-RSM, respectively. When the analysis model was the RSM, only one of the 19 estimators had a Z statistic beyond the ±2.576 critical range. When the analysis model was the RE-RSM, three estimators for the item parameters and all four estimators for the threshold variances were biased. However, their magnitudes of bias were very small, between –.020 and .065. Therefore, when the generating model was the RSM and the data were analyzed using the RE-RSM, the estimation bias was negligible and the estimated threshold variances were all very close to zero.

According to the likelihood ratio test, the difference in the two likelihood deviances ($G^2_{\text{RSM}} - G^2_{\text{RE-RSM}}$) would have to exceed the critical value of 9.49 ($df = 4$) to be statistically significant at the .05 nominal level. It turned out that the differences in the two likelihood deviances over the 100 replications were between –9.27 and 6.75. Therefore, these two models were not statistically significantly different. Note that even estimating four additional parameters in the RE-RSM, $G^2_{\text{RE-RSM}}$ was not always smaller than G^2_{RSM}. In theory, when more parameters are estimated, the likelihood deviance is expected to be smaller. However, this many not be true at the boundaries of parameter span, such as when a variance becomes zero, or when there are qualitative structural differences between models, such as when certain crucial parameters vanish (Stram & Lee, 1994, 1995). The negative differences in the likelihood deviances were more likely because estimating presumably zero variances in the RE-RSM solicited some difficulty in parameter estimation. In practice, we suggest that both the RSM and the RE-RSM should be used to fit the same data. When there is no statistical difference in the likelihood deviances (or the difference is negative), or the estimated variances for the thresholds in the RE-RSM are all very close to zero, Occam's Razor suggests that the RSM is preferred.

Regarding the item fit statistics, under the RSM, the outfit and infit MSE statistics had mean values of approximately 1.0, and the corresponding t statistics had mean values of approximately zero, and ranges approximately between ±3.0. As expected, the MSE and *t* statistics followed their theoretical distributions approximately, which means that the data fitted the RSM's expectation very well. Under the RE-RSM, the responses to the 15 item difficulties appeared to fit the model's expectation very well, too, in that the outfit and infit MSE statistics had mean values of approximately 1.0, and the corresponding *t* statistics had mean values of approximately zero, and ranges approximately between ±3.0. However, the four thresholds had mean outfit and infit MSE statistics around .95 (between .80 and 1.11), and mean outfit and infit *t* statistics around –.8 (between –3.36 and 1.72). Because the magnitudes of misfit were small, one can conclude that the RE-RSM also fitted the data fairly well.

Condition 2: Random Threshold Effects

In this condition, the generating model was the RE-RSM and the data were analyzed using the RSM and the RE-RSM, respectively. When the analysis model was the RSM, the *Z* statistics were between –20.15 and 31.75, a range far beyond the ±2.576 critical range. All the item difficulties were biased, all the threshold parameters were overexpanded, and the variance of θ was overestimated. In contrast, when the analysis model was the RE-RSM, all the estimators were unbiased, with *Z* statistics between –1.91 and 2.46. Good fit was expected because the RE-RSM was the generating model. The differences in the likelihood deviances between the two models ($G^2_{\text{RSM}} - G^2_{\text{RE-RSM}}$) ranged from 752.00 to 1204.63, respectively. Clearly, the RE-RSM fitted the data much better than the RSM.

With respect to the item fit statistics, for the 15 item difficulties under the RSM, the outfit and infit MSE statistics had mean values of approximately 1.0, and their corresponding *t* statistics had mean values of approximately zero and ranges approximately between ±3.0. On the other hand, for the three threshold parameters, the outfit and infit MSE statistics had mean values from approximately 2 to 3, and the corresponding *t* statistics had mean values from approximately 15 to 20. Consequently, the responses to the three thresholds did not fit the model's expectation. Under the RE-RSM, all the responses appeared to fit the model's expectation very well, in that the outfit and infit MSE statistics had mean values of approximately 1.0, and the corresponding *t* statistics had mean values of approximately zero, and ranges approximately between ±3.0. It appeared that a very large MSE or *t* for the threshold parameters in the RSM signified randomness in the thresholds and suggested the RE-RSM.

Two major conclusions can be drawn from the simulations. First, when the data were generated from the RSM, using the RSM and the RE-RSM to fit the data made little difference, in terms of the likelihood deviances, item fit statistics, and threshold variance estimates, and hence the simpler RSM model was preferred. Second, when the data were generated from the RE-RSM, using the RE-RSM to fit the data led to accurate parameter recovery, expected infit and outfit statistics, whereas using the RSM resulted in biased item difficulties, overexpanded thresholds, overestimated variance of θ and large infit and outfit statistics for the thresholds. Practitioners are thus recommended to fit both the RSM and the RE-RSM to rating scale data and compare them in terms of likelihood deviances, item fit statistics, and estimates for the threshold variances.

An Empirical Example

A total of 14,022 high school students in Taiwan were randomly sampled and asked to generate a response to 10 statements, which were ranked on four-point scale (Wu, Yin, Lee, & Hu, 2001). These statements included expressions such as "When someone insults me or my family, I would feel better if I can beat them up" and "If I do not beat the one who laughs at me, I would be upset." The respondents were asked to describe how each statement corresponded with their feelings: (1) does not match at all, (2) does not match, (3) matches to some extent, and (4) matches exactly. The higher the raw score, the greater the tendency toward violent aggression would be. For simplicity, we randomly selected 500 junior high school students. Of them, 49.6% were boys, 48.8% were girls, and 1.6% were missing; 31.6% were seventh graders, 42.0% were eighth graders, and 24.8% were ninth graders. Classical item analyses suggested that these 10 items constituted a single dominant dimension. For instance, the first five eigenvalues in the principal component analysis were 4.68, .76, .69, .65, and .63, respectively. Note that the RE-RSM and the CRE-PCM treat subjective judgments in rating scales as additional dimensions so that these two models are actually multidimensional. Classical principal component analysis or factor analysis may not be sensitive enough to detect this kind of multidimensionality. Fortunately, it is found from the previous simulations that fit statistics for the thresholds are very sensitive to it.

The RSM, the PCM, the RE-RSM, and the CRE-PCM were fitted to the data. Test reliability was computed as follows. Mislevy, Beaton, Kaplan, and Sheehan (1992) showed that when the latent trait θ is normally distributed, its population variance estimate is obtained from MML estimation, and the posterior distribution is approximately normally distributed, then test reliability can be computed as:

$$\text{Reliability} = \text{Var}(\hat{\theta}_{\text{EAP}})/\hat{\sigma}^2, \tag{18}$$

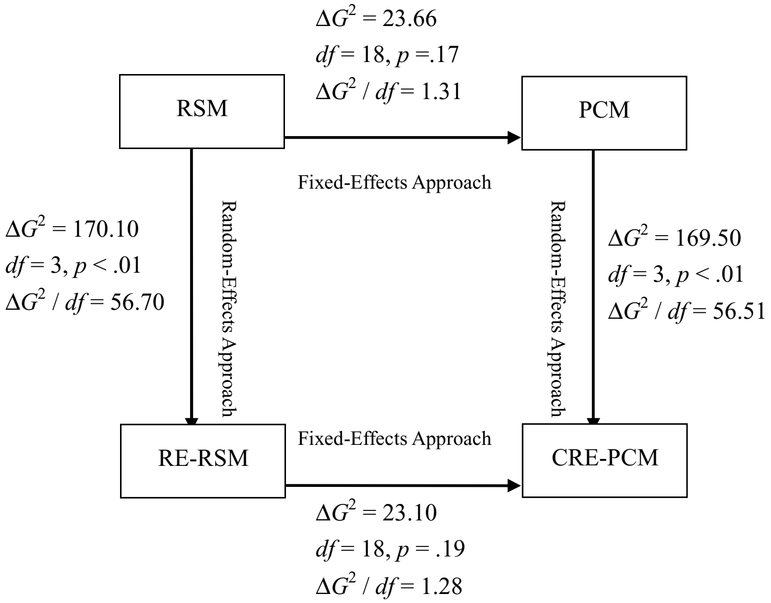
where $\hat{\sigma}^2$ is the population variance estimate of the latent trait θ and $\text{Var}(\hat{\theta}_{\text{EAP}})$ is the sampling variance of the EAP estimates over persons (each person has an EAP estimate). This test reliability, being a counterpart of classical test reliability, can be directly applied in this study, because the latent trait and the random threshold variables were all assumed to be normally distributed, MML estimation was used, and the posterior distributions were approximately normally distributed.

Table 1 lists the likelihood deviance (-2 loglikelihood), number of parameters, Akaike's information criterion, sampling variances of the EAP estimates, population variance estimates, and test reliability for the four models. The RE-RSM had the smallest AIC so that it was the best-fitting model. Figure 2 shows the likelihood ratio tests. The RSM did not differ statistically significantly from the PCM ($p = .17$), nor did the RE-RSM differ from the CRE-PCM ($p = .19$). Comparing the RSM with the PCM, and the RE-RSM with the CRE-PCM, one finds that adding one additional

TABLE 1
Likelihood Deviances, Numbers of Parameters, Akaike's Information Criterion, Sampling Variances of EAP Estimates, Variance Estimates, and Test Reliabilities in the Four Models

Model	Deviance	Number of Parameters	AIC	$\text{Var}(\hat{\theta}_{\text{EAP}})$	$\hat{\sigma}^2$	Reliability
RSM	7911.00	13	7937.00	1.71	2.10	.81
PCM	7887.34	31	7949.34	1.70	2.09	.81
RE-RSM	7740.90	16	7772.90	.44	.95	.46
CRE-PCM	7717.80	34	7785.80	.47	1.04	.45

Note. AIC = Deviance + two times the number of parameters; $\text{Var}(\hat{\theta}_{\text{EAP}})$: sampling variance of EAP estimates; $\hat{\sigma}^2$ = variance estimate.



Note. RSM = rating scale model; PCM = partial credit model; RE-RSM = random-effects rating scale model; CRE-PCM = constrained random-effects partial credit model.

FIGURE 2. *Likelihood ratio tests for the four models*

fixed-effects parameter reduced the likelihood deviances by only 1.31 and 1.28 on average, respectively. In contrast, comparing the RSM with the RE-RSM, and the PCM with the CRE-PCM, one finds that adding one additional random-effects parameter reduced the likelihood deviances by 56.70 and 56.51 on average, respectively. Thus, allowing each item to have its own set of rating scale structure under the PCM (the fixed-effects approach) did not improve model-data fit as remarkably as allowing the thresholds to be random under the RE-RSM (the random-effects approach). In short, the random-effects approach was much more efficient than the fixed-effects approach in capturing subjective judgments and improving model-data fit for the rating scale items.

TABLE 2

Parameter Estimates and Item Fit Statistics under the RE-RSM and the RSM

	RE-RSM					RSM				
Para	Est	Infit MSE	Infit t	Outfit MSE	Outfit t	Est	Infit MSE	Infit t	Outfit MSE	Outfit t
σ_θ^2	.95					2.10				
$\sigma_{\gamma_1}^2$	2.18									
$\sigma_{\gamma_2}^2$	.83									
$\sigma_{\gamma_3}^2$	1.40									
β_1	.99	1.03	.5	1.08	1.3	1.18	1.23	3.2	1.41	5.8
β_2	2.04	.95	−.5	1.02	.3	2.26	.94	−.8	.96	−.6
β_3	2.16	.93	−.7	1.35	5.0	2.37	.90	−1.2	.98	−.4
β_4	2.26	.93	−.6	.99	−.2	2.47	.99	−.1	.82	−3.0
β_5	2.16	.91	−.9	.88	−2.0	2.37	.97	−.4	.92	−1.2
β_6	1.69	1.06	.7	1.06	.9	1.90	1.09	1.2	1.02	.3
β_7	2.34	1.04	.4	1.15	2.3	2.55	.96	−.4	.89	−1.8
β_8	1.99	.95	−.5	1.01	.2	2.20	.97	−.4	1.03	.5
β_9	.88	1.02	.4	.93	−1.1	1.05	1.07	1.1	1.00	.0
β_{10}	2.21	1.12	1.2	.96	−.6	2.42	1.09	1.1	1.17	2.6
τ_1	−.71	1.03	.4	1.03	.4	−1.05	1.76	9.3	1.53	7.3
τ_2	.30	.90	−1.0	.81	−3.3	.44	1.45	4.1	1.94	11.8
τ_3	.41	.97	−.2	.96	−.5	.61	1.76	4.8	3.08	21.6

Note: Para = Parameter; Est = Estimate; MSE = Mean Square Error; σ_θ^2 denotes the variance of θ; $\sigma_{\gamma_1}^2$ through $\sigma_{\gamma_3}^2$ are the variance estimates of the three random thresholds, respectively; β_i denotes the overall difficulty of item i; τ_j denotes the j-th threshold parameter.

The parameter estimates and fit statistics for the item difficulties and the thresholds under the RE-RSM and the RSM are shown in Table 2. The mean item difficulties for the 10 items were 1.87 and 2.08 under the RE-RSM and the RSM, respectively. Because the mean of the latent trait was set at zero for model identification, a mean item difficulty of 1.87 or 2.08 indicated that these 10 items were very difficult for these persons to endorse: They were not very violent because they were representative samples from the normal population rather than from the young offenders. Compared to the RE-RSM, the RSM overexpanded the thresholds and overestimated the variance of θ. The patterns of biased estimation in the RSM were consistent with those found in the previous simulations.

According to the item fit statistics, the responses to the item difficulties fitted both models' expectation fairly well. The responses to the three threshold parameters also

TABLE 3

Descriptive Statistics of the Standard Errors of the EAP Estimates under the RSM and the RE-RSM

Model	Mean	SD	Minimum	Maximum
RSM	.61	.22	.98	.34
RE-RSM	.71	.07	.80	.54

fitted the RE-RSM's expectation fairly well, with fit MSE statistics between .81 and 1.03 and *t* statistics between –3.3 and .4. In contrast, they fitted the RSM's expectation very poorly, with fit MSE between 1.45 and 3.08 and *t* statistics between 4.1 and 21.6.

Under the RE-RSM, the point estimates for the three thresholds were –.71, .30, and .41 (they summed to zero for model identification), respectively, and the variance estimates were 2.18, .83, and 1.40, respectively. In other words, $\tau_{n1} \sim N(-.71, 2.18)$, $\tau_{n2} \sim N(.30, .83)$, and $\tau_{n3} \sim N(.41, 1.40)$. Because the variances were rather large, a great proportion in test score variation was attributable to the randomness in the thresholds over persons.

Equation (18) as well as classical test reliability is actually an index of sample separation. If the sample happens to be homogeneous (no individual differences), the test reliability will be zero. If the sample happens to be extremely heterogeneous, the test reliability will be close to unity. In this sense, test reliability is sample dependent. In contrast, standard error of measurement, which is computed from test information function, does not depend on samples (Lord, 1980). Therefore, in addition to the test reliability, it is useful to compare the distributions of the standard error of person measures under the RSM and the RE-RSM. Table 3 lists the means, standard deviations, minimum and maximum values of the standard error of the person EAP estimates (i.e., standard deviation of the posterior distribution). The mean standard errors for the RSM and the RE-RSM were similar, .61 and .71, respectively, suggesting that these two models yielded similar degrees of measurement precision. Because the estimate for the variance of θ under the RE-RSM was small (.95), the test reliability was only .46. In contrast, the variance of θ under the RSM was overestimated (2.10), so was the test reliability (.81). The overestimation for the variance of θ in the RSM was because the randomness in the thresholds was not correctly taken into account so that it was added into the variance of θ. Under the RE-RSM, the randomness was directly considered so that the estimates for the variance of θ and the test reliability were more appropriate.

Conclusion and Discussion

When responding to rating scale items, a person generally relies on subjective judgment to select a category that best describes his or her situation. Different persons may use different criteria in making the judgment. To describe the variation in subjective judgment between persons, we add random variables, one for each threshold, into the RSM. The resulting RE-RSM turns out to be a special case of the MRCMLM, so that the estimation procedure available in ConQuest can be directly used. Through simulation, it appears that when the data are generated from the RSM, using the RSM and the RE-RSM to fit the data makes little difference: These two models yield statistically equivalent model-data fit, and the RE-RSM yields near zero variance estimates for the random thresholds. When the data are generated from the RE-RSM, using the RE-RSM to fit the data leads to unbiased estimates, whereas using the RSM yields biased item difficulties, overexpanded thresholds, overestimated variance, and large fit statistics for the thresholds.

Note that in these simulations, the mean person latent trait was set to be identical to the mean item difficulty (i.e., the person-item match is perfect). In practice, persons may have latent trait levels far above or below the mean item difficulty (e.g., as in the empirical example). Whether the findings from the simulations can be generalized to situations where the person-item match is poor needs further investigation.

The empirical example shows that the RE-RSM fits the data of the 10 four-point rating scale items much better than the RSM and the PCM. No statistically significant difference between the RE-RSM and the CRE-PCM is found. The variance estimates for the three random thresholds obtained from the RE-RSM are fairly large, suggesting that the randomness in the thresholds is substantial. As a result, there is great amount of randomness between persons in judging the four points. The test reliability is only .46 when the randomness is correctly taken into account. When the randomness is ignored under the RSM, the test reliability is overestimated to be .81. According to standard error of measurement, the RE-RSM yields similar degree of measurement precision as the RSM. As in the simulations, the RSM yields biased item difficulties, overexpanded thresholds, overestimated variance, and large fit statistics for the thresholds.

In practice, large fit statistics for the thresholds in the RSM may suggest randomness in the thresholds and thus call for the RE-RSM. If the variance estimates for the thresholds in the RE-RSM are large, and the difference in the likelihood deviances between the RSM and the RE-RSM is large, one may conclude that the RE-RSM captures the randomness in the thresholds better than the RSM. On the other hand, if the variance estimates for the thresholds in the RE-RSM are all very close to zero, and the difference in the likelihood deviances between the RSM and the RE-RSM models is very small, then the RSM and the RE-RSM show little practical difference and thus the RSM is better.

In this study, the threshold effects are treated as "random noises" and are assumed to be uncorrelated with each other and with the latent trait. That is, the structure of the random noises and the latent trait is orthogonal. This orthogonal structure is analogous to the standard situation in exploratory factor analysis, in which common factors are assumed to be independent, and unique factors are assumed to be independent with each other and with common factors. In fact, the orthogonal structure is not necessary for identification of the parameters in the RE-RSM. Under certain conditions (e.g., where one of the thresholds is constrained to have no randomness), the structure can be relaxed to be nonorthogonal. The orthogonal structure is imposed in this study for several reasons. First, the interpretation of the latent trait is simpler and more direct in the orthogonal structure than in the nonorthogonal structure. When the random noises are correlated with each other and with the latent trait, the latent trait may be seriously contaminated so that it is not clear what the resulting measures reflect. Second, in order to implement the nonorthogonal structure, one needs to allocate one threshold as fixed, which is difficult and arbitrary to decide. Third, the orthogonal structure has been commonly imposed in other similar random-effects models (e.g., Bradlow et al., 1999; Wainer et al., 2000; Wang et al., 2002; Wang & Wilson, 2005a, 2005b). Finally, computation time is substantially reduced and the iterations converge more easily, compared to what happens in computations for the nonorthogonal structure. Allowing the random noises to be correlated with each other and with the latent trait may increase model-data fit to some extent, although possibly at the expense of model complexity and threats to test validity. It is left for future studies to compare these two kinds of structure.

Although rating scale items are taken as an example in this study, the RE-RSM can be applied to other types of items, as long as randomness in the thresholds is suspected. For example, randomness may occur when a rater gives ratings to persons' performances (i.e., intra-rater effects) (Wang & Wilson, 2005a). Another example would be ability items that are scored according to a common framework, such as the Structure of the Observed Learning Outcome (SOLO) taxonomy (Biggs & Collis, 1982), which provides a systematic way of describing how a learner's performance grows in complexity when mastering tasks through several stages: prestructural, unistructural, multistructural, relational, and extended abstract. Standard Rasch analyses have been applied to analyze items that are scored using the SOLO taxonomy (Scholten, Keeves, & Lawson, 2002; Wilson, 1992; Wilson & Iventosch, 1988; Wilson & Sloane, 2000). As the items are scored by raters according to common scoring rubrics, variations in the thresholds can occur. It would be interesting to check magnitudes of the variations by fitting the RE-RSM to the data.

The RE-RSM, although being an extension of the RSM, does not possess the same desirable measurement properties of the RSM; for example, person raw scores alone are no longer sufficient statistics of person measures. Instead, the vector of scores on all dimensions (θ and γ) is needed. The RE-RSM serves two main purposes: scaling and screening. For scaling purposes, the RE-RSM can be used to detect randomness in the thresholds over persons, from whom test analysts acquire information about the magnitudes of randomness in the thresholds, individual differences on the latent trait and the random threshold variables, and more accurate estimates for test reliability. For screening purposes, if the magnitudes of randomness in the thresholds are large and the test reliability is low, practitioners are encouraged to explore the causes and try to reduce the magnitudes so as to build a better rating scale, to achieve a higher test reliability, and to reach a higher measurement quality. This is certainly not an easy job. A starting point might be to regress the random threshold variables on personal background variables (e.g., gender, age, education level, and socio-economic status) or item characteristics. Vague and abstract statements for items or rating scale labels might solicit more randomness than behavioral-based and concrete statements. Young children and persons with low education levels might show more randomness. Further empirical studies are needed to investigate the causes of randomness and how to reduce it.

As the RSM can be extended to the RE-RSM so can be the PCM to the RE-PCM or the CRE-PCM. Investigation and application of the RE-PCM and the CRE-PCM are left for future studies. The RE-RSM, the RE-PCM, and the CRE-PCM belong to the family of one-parameter Rasch models. Future studies may generalize the random-effects approach to two-parameter multi-category item response models, such as the graded response model (Samejima, 1969), the nominal response model (Bock, 1972), and the generalized partial credit model (Muraki, 1992).

References

Adams, R. J., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1–23.

Adams, R. J., Wilson, M. R., & Wu, M. (1997). Multilevel item response models: An approach to errors in variable regression. *Journal of Educational and Behavioral Statistics, 22,* 47–76.

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

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

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

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

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431–444.

Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

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

De Boeck, P., & Wilson, M. R. (Eds.) (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

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

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 269–286.

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

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

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

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133–161.

Mislevy, R. J., & Bock, R. D. (1989). A hierarchical item-response model for educational testing. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 57–74). San Diego, CA: Academic Press.

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

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A, 135,* 370–384.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* 342–366.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogische Institut.

Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement, 26,* 269–283.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185–205.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53,* 349–359.

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

SAS Institute (1999). *The NLMIXED procedure* [Computer software]. Cary, NC: Author.

Scholten, I., Keeves, J. P., & Lawson, M. J. (2002). Validation of a free response test of deep learning about the normal swallowing process. *Higher Education, 44,* 233–255.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247.

Smith, E. V., Jr., & Smith, R. M. (Eds.) (2004). *Introduction to Rasch Measurement*. Maple Grove, MN: JAM Press.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed-effects model. *Biometrics, 50,* 1171–1177.

Stram, D. O., & Lee, J. W. (1995). Correction to: Variance components testing in the longitudinal mixed-effects model. *Biometrics, 51,* 1196.

Volodin, N., & Adams, R. J. (1995). *Identifying and estimating a D-dimensional Rasch model.* Paper presented at the International Objective Measurement Workshop. University of California at Berkeley, California.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8,* 157–186.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–202.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement, 57,* 749–766.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22–29.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37,* 203–220.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). London: Kluwer.

Wang, W.-C., & Wilson, M. R. (2005a). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29,* 296–318.

Wang, W.-C., & Wilson, M. R. (2005b). The Rasch testlet model. *Applied Psychological Measurement, 29,* 126–149.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26,* 109–128.

Wilson, M. R. (1992). The partial order model: An extension of the partial credit model. *Applied Psychological Measurement, 16,* 309–425.

Wilson, M. R., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika, 60,* 181–208.

Wilson, M. R., & Iventosch, L. (1988). Using the partial credit model to investigate responses to structured subtests. *Applied Measurement in Education, 1,* 319–334.

Wilson, M. R., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13,* 181–208.

Wolfinger, R. D., & SAS Institute (n.d.). Fitting nonlinear mixed models with the new NLMIXED procedure. Retrieved August 17, 2003, from http://support.sas.com/rnd/app/papers/nlmixedsugi.pdf.

Wright, B. D. (1999). Model selection: Rating scale or partial credit? *Rasch Measurement Transactions, 12*(3), 641–642.

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

Wright, B. D., Linacre, J. M., Gustafson, J.-E., & Martin-Löf, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wu, J., Yin, M.-C., Lee, K.-C., & Hu, S.-C. (2001). *Prevention and control of school violence in Taiwan*. (NSC Research Report No. NSC89-2420-H006-001-QBS). Taipei, Taiwan: The National Science Council.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest* [Computer software and manual]. Camberwell, Victoria, Australia: Australian Council for Educational Research.

Yen, W. (1993). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–213.

Authors

WEN-CHUNG WANG is a Professor, Department of Psychology, National Chung Cheng University, Chia-Yi, 621, Taiwan; psywcw@ccu.edu.tw. His primary research interests include psychometrics.

MARK WILSON is a Professor, Graduate School of Education, University of California at Berkeley, 4415 Tolman Hall, Berkeley, CA 94720; markw@calmail.berkeley.edu. His primary research interests include educational measurement.

CHING-LIN SHIH is a Doctoral Candidate, Department of Psychology, National Chung Cheng University, Chia-Yi, 621, Taiwan; scl@mail.ntcu.edu.tw. His primary research interests include psychometrics.