Item Response Theory Models for Carry-Over Effect Across Different Scales

Applied Psychological Measurement
2015, Vol. 39(5) 406–425
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621615572250
apm.sagepub.com

Kuan-Yu Jin¹ and Wen-Chung Wang¹

Abstract

It is common in educational and psychological tests or social surveys that the same statement is judged on multiple scales. These multiple responses are linked by the same statement, which may cause local dependence. Considering the way a statement is judged on multiple scales, a new class of item response theory (IRT) models is developed to account for the nonrecursive carry-over effect, in which a response can be affected only by its preceding response rather than by a subsequent response. The parameters of the models can be estimated with the freeware WINBUGS. Two simulation studies were conducted to evaluate the parameter recovery of the new models and the consequences of model misspecification. Results showed that the parameters of the new models were recovered fairly well; fitting unnecessarily complicated models to data that did not have the carry-over effect did little harm to parameter estimation; and ignoring the carry-over effect by fitting standard IRT models yielded biased estimates for the item parameters, the correlation between latent traits, and the test reliability. Two empirical examples with parallel design and sequential design are provided to demonstrate the implications and applications of the new models.

Keywords

local item dependence, parallel design, sequential design, item response theory, Bayesian methods

It is common that a statement is judged across several scales. For example, a symptom is judged according to frequency and severity, which is referred to as the Presence–Severity (P–S) format (Liu & Verkuilen, 2013); a 21st-century core competency is judged according to procession and importance; and a commercial product is judged on price and quality. Hereafter, this kind of tests is called one-statement-multiple-scale (OSMS) instruments. There are two major arrangements in OSMS instruments: parallel design and sequential design (Wang, Cheng, & Wilson, 2005). In a typical parallel design, statements are shown (printed) on the left-hand side of a page, and scales are printed in parallel on the right-hand side, which enables persons to respond to a statement according to these scales, and then move to the next statement, and so on. In a typical sequential design, all statements together with the first scale are printed in one section,

¹The Hong Kong Institute of Education, New Territories, Hong Kong

Corresponding Author:

Wen-Chung Wang, Department of Psychological Studies, The Hong Kong Institute of Education, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong.
Email: wcwang@ied.edu.hk

and the same statements together with another scale in another section, and so on. This sequential design makes respondents respond to the statement scales on one scale, followed by the same statements on another scale. When statements are judged according to multiple scales in a parallel or sequential design, a common practice in data analysis is to treat each scale (in conjunction with the statements) as a distinct test measuring a distinct latent trait. For instance, if an OSMS instrument consists of 10 statements and 2 scales, then they will be treated as 2 distinct tests, each with 10 items.

In the framework of item response theory (IRT), a common practice of analyzing an OSMS instrument is to fit an IRT model to each of the tests separately. In doing so, it is assumed that item residuals are mutually independent after the latent traits are controlled. This is referred to as the assumption of local independence, which enables the use of products of likelihoods for parameter estimation. Responses to a statement on multiple scales may not be locally independent because these “items in different tests” are linked by the same statement. For instance, an endorsement on a high frequency of suffering a symptom may affect the next endorsement on the severity of the same symptom. This situation is analogous to the case in which a correct answer on an item increases the probability of success on another item. If so, items are locally dependent.

When items are locally dependent, standard IRT models that fail to account for such dependence are no longer appropriate. Fitting a wrong IRT model will produce incorrect likelihoods, which in turn will yield incorrect parameter estimates. The major purpose of this study is to propose a new class of IRT models to account for the carry-over effect in OSMS instruments.

There are two major kinds of local item dependence (LID; Chen & Thissen, 1997): underlying local item dependence (ULID) and surface local item dependence (SLID). ULID suggests that a set of items share a common latent trait, which is irrelevant to the intended-to-be-measured latent trait. Ignoring ULID by fitting standard IRT models will overestimate test reliability and yield biased parameter estimates. Testlet and bi-factor IRT models (Cai, 2010; Li, Bolt, & Fu, 2006; Wainer, Bradlow, & Wang, 2007; Wang & Wilson, 2005) have been developed to account for ULID, in which a specific latent trait that is independent of the target latent trait is added to each testlet. SLID refers to a condition where LID occurs because items are highly similar in contents or locations. When SLID takes place, the parameters of an item will be affected by its preceding (or similar) items (Yousfi & Böhme, 2012). In OSMS instruments, the same statement is responded multiple times along different scales, which can be viewed as an example of SLID.

Existing IRT Models for Local Item Dependence

In the literature on IRT models for LID, most studies focus on accounting for LID within tests. For example, consider the Neuropsychiatry Inventory (NPI) Scale, in which 10 neuropsychiatric disturbances that occur frequently in dementia are rated on a 4-point (1-4) frequency scale and a 3-point (1-3) severity scale (Cummings et al., 1994). These two scores are multiplied so that each neuropsychiatric disturbance would get one of these scores: 1, 2, 3, 4, 6, 8, 9, and 12. Within the IRT framework, one can then fit IRT models to these 8-point “composite” items, for example, the (generalized) partial credit model (Masters, 1982; Muraki, 1992). A drawback of this approach is that these eight categories may not reflect the same latent trait, and the ordinal nature may not hold, so unidimensional ordinal IRT models may not apply. To account for the nonordinal nature, one may adopt the nominal response model (Bock, 1972), in which these eight categories are treated as nominal data. In fact, the use of the nominal response model is not without problem. Note that in the formulation of new scores, for example, a score of 3 would have two combinations. It can be generated from 1 on the frequency scale and 3 on the

severity scale, or 3 on the frequency scale and 1 on the severity scale. These two combinations are merged into a score of 3, but they have very different meanings.

One way to account for the fine difference between the combinations of (1, 3) and (3, 1) is to create a total of 12 categories (response patterns) rather than 8 categories. That is, there are (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3), (4, 1), (4, 2), and (4, 3). IRT models can be fit to these 12-category composite items, which can be unidimensional models (Hoskens & Boeck, 1995; Tuerlinckx & De Boeck, 1999; Wilson & Adams, 1995) or multidimensional models (Hoskens & De Boeck, 2001). To be in line with the intension of test developers, that different scales in an OSMS instrument are designed to measure different latent constructs, multidimensional IRT models are preferred.

Hoskens and De Boeck (2001) present a multidimensional IRT model to account for LID in cognitive tests. For illustrative simplicity, let there be two dichotomous items and a total of four response patterns: (0, 0), (1, 0), (0, 1), and (1, 1), where 0 and 1 are the scores of the two dichotomous items. Their probabilities are denoted as $P_{ns(1,0)}$, $P_{ns(0,1)}$, $P_{ns(0,0)}$, and $P_{ns(1,1)}$, respectively, where s indexes the item bundles, and n indexes the persons. The relationships among the four patterns are as follows:

$$\log[P_{ns(1,0)}/P_{ns(0,0)}] = \theta_{n1} - \delta_{s1}, \quad (1)$$

$$\log[P_{ns(0,1)}/P_{ns(0,0)}] = \theta_{n2} - \delta_{s2}, \quad (2)$$

$$\log[P_{ns(1,1)}/P_{ns(0,0)}] = (\theta_{n1} - \delta_{s1}) + (\theta_{n2} - \delta_{s2}) - \tau_s, \quad (3)$$

where θ_{n1} and θ_{n2} are the two latent traits of Person n , δ_{s1} and δ_{s2} are the difficulty parameters of the two items within Bundle (composite-item) s , and τ_s describes the interaction between the two items. A negative τ_s increases the probability of scoring 1 on both items; a positive τ_s decreases the probability of scoring 1 on both items. When τ_s is zero, these two items are locally independent. Although this bundle approach appears promising, it suffers from a large number of response patterns. For instance, if the two items are polytomous, and each has five categories, then there will be 25 response patterns; if there are five dichotomous items, then there will be 32 response patterns. In such cases, a large number of the τ parameters will be needed to describe the LID within each bundle. Unless the sample size is very large, this bundle approach is not feasible when the number of response patterns is large. Furthermore, when there are many τ parameters within a bundle, the interpretation of these τ parameters will be very complicated.

Adopting the same logic, Wang et al. (2005) propose a multidimensional IRT model to describe the LID in OSMS instruments. For illustrative simplicity, let statement (indexed s) be judged on two scales. The log-odds of the two scores for Statement s are modeled as follows:

$$\log[P_{nsk1}/P_{ns(k-1)1}] = \theta_{n1} - \delta_{sk1} - \tau_{sm}, \quad (4)$$

$$\log[P_{nsk2}/P_{ns(k-1)2}] = \theta_{n2} - \delta_{sk2} - \tau_{sm}, \quad (5)$$

where P_{nsk1} and $P_{ns(k-1)1}$ are the probabilities of receiving scores of k and $k-1$ for Statement s on Scale 1, respectively; P_{nsk2} and $P_{ns(k-1)2}$ are the probabilities of receiving scores of k and $k-1$ for Statement s on Scale 2, respectively; θ_{n1} and θ_{n2} are the two latent traits of Person n ; δ_{sk1} and δ_{sk2} are the k th step difficulties of the two items for Statement s ; τ_{sm} ($m = 1, \dots, (I-1) \times (J-1)$; I and J are the numbers of response categories of the two items) describes the LID and is referred to the dependency parameter. When τ_{sm} is constrained to be identical across all statements, the following model is formed:

$$\log[P_{nsk1}/P_{ns(k-1)1}] = \theta_{n1} - \delta_{sk1} - \tau_m, \quad (6)$$

$$\log[P_{nsk2}/P_{ns(k-1)2}] = \theta_{n2} - \delta_{sk2} - \tau_m. \quad (7)$$

If all the τ parameters are zero, items are actually locally independent, and standard IRT models such as the partial credit model become feasible for each test.

Although Equations 4 to 7 appear promising in accounting for the LID in OSMS instruments, there are limitations. First, the units of the τ parameters are not comparable between latent traits. A value of 1 in τ has different effects on different latent traits because these latent traits often have different variances. Second, the τ parameters are incorporated to describe the magnitudes of LID, which is recursive, rather than the magnitude of carry-over effect, which, by definition, is nonrecursive.

It is possible to add a random effect to standard IRT models to account for LID. Testlet and bi-factor IRT models are two examples and have been commonly applied to account for LID among items within testlets (Cai, 2010; Li, et al, 2006; Wainer et al., 2007; Wang & Wilson, 2005). The major difference between testlet and bi-factor models is that the two latent traits that an item measures share a common slope in testlet IRT models, whereas they have two different slopes in bi-factor IRT models. For example, under the Rasch testlet model, the log-odds of a correct answer over an incorrect answer to Item i within Testlet d for a person with latent trait level θ_n are defined as follows:

$$\log[P_{nik}/P_{ni(k-1)}] = \theta_n - \delta_{ik} + \lambda_{nd(i)}, \quad (8)$$

where θ_n is the latent trait of Person n , δ_{ik} is the k th step difficulty parameter of Item i , and $\lambda_{nd(i)}$ is a random effect for Person n on Item i within Testlet d , which describes the interaction between Person n and Item i within Testlet d . $\lambda_{nd(i)}$ is independent of θ_n and assumed to follow a normal distribution across persons: $\lambda_{nd(i)} \sim N(0, \sigma_d^2)$, and σ_d^2 depicts the amount of LID effect in Testlet d . When a statement is judged on two scales, these two item responses connected by the same statement can be treated as a testlet, and Equation 8 can be revised as follows:

$$\log[P_{nsk1}/P_{ns(k-1)1}] = \theta_{n1} - \delta_{sk1} + \lambda_{ns}, \quad (9)$$

$$\log[P_{nsk2}/P_{ns(k-1)2}] = \theta_{n2} - \delta_{sk2} + \lambda_{ns}, \quad (10)$$

where θ_{n1} and θ_{n2} are the two latent traits of Person n , δ_{sk1} and δ_{sk2} are the k th step difficulties for the two items for Statement s , and λ_{ns} describes the interaction between Person n and Statement s , and is assumed to follow a normal distribution across persons: $\lambda_{ns} \sim N(0, \sigma_s^2)$.

This testlet (random effect) approach (Equations 9 and 10) shares the same limitations as those of Wang et al.'s (2005) approach (Equations 4-7) in that the units of the λ parameters are not comparable between the two latent traits, and they do not consider the nonrecursive nature of the carry-over effect in OSMS instruments. In practice, there are a few scales, mostly only two. When there are two scales, each testlet consists of two items. Using two items to estimate a specific latent trait (λ_s) in conjunction with two latent traits (θ_1 and θ_2) is seldom feasible. Another difficulty of this random-effect approach is high dimensionality. The number of dimensions is equal to the number of statements plus the number of scales. In practice, there are many statements to be rated.

In the aforementioned equations, the LID is treated as recursive. Sometimes, a response can be affected only by its preceding item, rather than by its subsequent item. IRT models have been developed to account for such a carry-over effect (Andrich, Humphry, & Marais, 2012; Andrich

& Kreiner, 2010; Marais & Andrich, 2008). Let Item i precede Item j and thus Item i affect Item j . Andrich et al. (2012) propose the following Rasch model to account for the carry-over effect:

$$\log[P_{nj k} / P_{nj(k-1)}] = \theta_n - \delta_{jk}^*, \quad (11)$$

$$\delta_{jk}^* = \delta_{jk} - d \text{ if } k \leq x_i, \quad (12)$$

$$\delta_{jk}^* = \delta_{jk} + d \text{ if } k > x_i, \quad (13)$$

where $P_{nj k}$ and $P_{nj(k-1)}$ are the probabilities of scoring k and $k-1$ on Item j for Person n , θ_n is the latent trait of Person n , δ_{jk}^* is the k th variant threshold of Item j depending on the response to the preceding Item i , and x_i is the score of Item i . The value of d represents the magnitude of dependence. If d is zero, then there is no carry-over effect; if d is positive, then thresholds δ_{jk} ($k < x_i$) shift to the left, whereas thresholds δ_{jk} ($k > x_i + 1$) shift to the right; if d is negative, then thresholds δ_{jk} ($k < x_i$) shift to the right, whereas thresholds δ_{jk} ($k > x_i + 1$) shift to the left. The model is unidimensional because there is a single random-effect variable θ . In addition, it is implicitly assumed that Items i and j have the same number of categories. The model is not applicable for OSMS instruments because it is unidimensional, and it does not allow for different numbers of response categories in different scales.

The New Class of IRT Models

In responding to OSMS instruments, it is very likely that an endorsement on a statement on a scale affects the next endorsement on the same statement on the next scale. By definition, such a carry-over effect is directional and nonrecursive. For example, an endorsement on a symptom along the presence scale may affect the succeeding endorsement on the same symptom along the severity scale. It is not justifiable that a succeeding endorsement along the severity scale affects its preceding endorsement on the same symptom along the presence scale.

Acknowledging that the carry-over effect in OSMS instruments is nonrecursive, the following class of IRT models is proposed. For illustrative simplicity, let there be two scales. Later on, the model shall be extended to more scales. For items along the first scale, by definition, there is no carry-over effect, and thus, the standard IRT models apply. For example, one can fit the partial credit model:

$$\log[P_{nsk1} / P_{ns(k-1)1}] = \theta_{n1} - \delta_{sk1}, \quad (14)$$

where P_{nsk1} and $P_{ns(k-1)1}$ are the probabilities of receiving scores of k and $k-1$ for Statement s of Scale 1, respectively; θ_{n1} is the first latent trait of Person n ; and δ_{sk1} is the k th step difficulties for Statement s of Scale 1. For Statement s of Scale 2, which may suffer from the carry-over effect from Statement s of Scale 1, the following is fit:

$$\log[P_{nsk2} / P_{ns(k-1)2}] = \theta_{n2} - \delta_{sk2} + \eta_{si1}, \quad (15)$$

where η_{si1} ($i = 0, \dots, I_{s1} - 1$; I_{s1} is the category number Statement s of Scale 1) presents the carry-over effect from Statement s of Scale 1 to that of Scale 2; the others are defined previously. η_{s01} needs to be constrained at zero as a reference. A positive η_{si1} increases the logit in Equation 15 (i.e., a positive carry-over effect), whereas a negative η_{si1} decreases the logit (i.e., a negative carry-over effect); a zero η_{si1} indicates no carry-over effect. When η_{si1} is zero for all i and s , Equation 15 is equivalent to the partial credit model, indicating that there is no carry-over effect in the OSMS instrument.

Suppose there are 10 statements and 2 scales in an OSMS instrument, the total number of the η parameters will be $\sum_{s=1}^{10} (I_{s1} - 1)$. If I_{s1} is 5 for all the 10 statements (i.e., the first scale has 5 points), then the total number will be 40. When appropriate, these 40 parameters can be constrained to form simpler models as follows:

$$\eta_{si1} = \eta_{s1}, \quad (16)$$

$$\eta_{si1} = \eta_{i1}, \quad (17)$$

$$\eta_{si1} = x_{s1} \eta_{s1}, \quad (18)$$

where Equation 16 assumes that the carry-over effect is identical across scores, so that each statement has 1 carry-over-effect parameter (i.e., there are 10 carry-over-effect parameters for the OSMS instrument); Equation 17 assumes that the carry-over effect is identical across statements, so that all statements share the same set of carry-over-effect parameters (i.e., there are 4 carry-over-effect parameters for the OSMS instrument); Equation 19 assumes that the carry-over effect is a linear function of the item scores (x_{s1} is the score of the first item of Statement s); and there are 10 carry-over-effect parameters for the OSMS instrument.

When there is a third scale, the endorsements on the third scale can be assumed to be affected by those on the preceding scale (i.e., the second scale). Based on this assumption, one can fit the following model to the responses on the third scale:

$$\log[P_{nsk3}/P_{ns(k-1)3}] = \theta_{n3} - \delta_{sk3} + \eta_{si2}, \quad (19)$$

where P_{nsk3} and $P_{ns(k-1)3}$ are the probabilities of receiving scores of k and $k-1$ for Statement s of Scale 3, respectively; θ_{n3} is the first latent trait of Person n ; δ_{sk3} is the k th step difficulties of Statement s of Scale 3; η_{si2} ($i = 0, \dots, I_{s2}-1$; I_{s2} is the number of category of Statement s of Scale 2) presents the carry-over effect from Statement s of Scale 2 onto that of Scale 3. The generalization to more than three scales is straightforward. When appropriate, item slope parameters can be incorporated, just like the two-parameter logistic model:

$$\log[P_{nsk1}/P_{ns(k-1)1}] = \alpha_{s1}(\theta_{n1} - \delta_{sk1}), \quad (20)$$

$$\log[P_{nsk2}/P_{ns(k-1)2}] = \alpha_{s2}(\theta_{n2} - \delta_{sk2} + \eta_{si1}), \quad (21)$$

$$\log[P_{nsk3}/P_{ns(k-1)3}] = \alpha_{s3}(\theta_{n3} - \delta_{sk3} + \eta_{si2}), \quad (22)$$

where α_{s1} , α_{s2} , and α_{s3} are the slope parameters for Statement s of Scales 1, 2, and 3, respectively. Equations 14 to 22 are referred to as the multidimensional model for carry-over effect in OSMS (MMCEO) instruments.

Parameter Estimation

The parameters in the MMCEO can be estimated with marginal maximum likelihood estimation methods, which have been implemented on computer programs such as SAS NLMIXED (SAS Institute, 1999), or the Bayesian estimation with Markov chain Monte Carlo (MCMC) methods, which have been implemented on computer programs such as the freeware WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2007). WinBUGS is used in this study because it is free and flexible. In Bayesian estimation, a statistical model and prior distributions of model parameters are specified to yield a joint posterior distribution. Because the joint posterior

distribution is often very difficult to obtain, MCMC methods are used. After a sequential sampling, the posterior distribution of each parameter is obtained. Its mean and standard deviation are reported as a point estimate and corresponding standard error (*SE*) of the estimate.

In the following simulation study and empirical example, two scales are focused on. To identify the MMCEO, the latent traits are assumed to follow a bivariate normal distribution with means of zero and variances of one.

Simulation Studies

Design and Analysis

Two studies were conducted. Study 1 focused on the recovery of the MMCEO. Item responses were generated from the MMCEO with the linear constraint of Equation 18, denoted as MMCEO-L. Three models were fit to the simulated data. One was the data-generating MMCEO-L, one was the MMCEO (no constraint on the η parameters), and the other was the generalized partial credit model (GPCM), in which the carry-over effect was ignored. All these models had item slope parameters. There were 2,000 persons, 10 statements, and two 4-point scales. The independent variables were (a) the magnitude of η_s : 0 and $-.3$ for all statements, and (b) the correlation between latent traits ρ : 0, $.4$, and $.8$. A positive (negative) η_s indicated that respondents who endorsed a high score on Statement s of Scale 1 tended to exaggerate (suppress) their attitudes toward the same statement of Scale 2. The magnitude of $.3$ was similar to those found in the two following empirical examples. A ρ of 0, $.4$, and $.8$ denoted 0, small, and large associations, respectively. The mean of the two latent traits was 0, and the variance of the two latent traits was 1. The slope parameters (α_{s1} and α_{s2}) were generated from lognormal ($0, .3^2$), and the item thresholds (δ_{sk1} and δ_{sk2}) were set between -2 and 2 . There were 100 replications in each condition. The outcome variables included the bias and root mean square error (RMSE) in the parameter estimates.

It was expected that fitting the data-generating MMCEO-L would yield good parameter recovery; fitting the unnecessarily complicated MMCEO would do little harm on parameter recovery; ignoring the carry-over effects by fitting the GPCM would yield poor parameter recovery. Note that when $\eta_s = 0$, there was no carry-over effect, so the GPCM was actually the data-generating model and the MMCEO-L and MMCEO became unnecessarily complicated models. Even so, it was expected that there would be little harm in parameter recovery for the MMCEO-L and MMCEO.

When η_s was negative (e.g., $-.3$) but ignored by fitting the GPCM, the estimates for the δ parameters of Scale 2 would not be biased for those respondents who endorsed 0 on the same statements of Scale 1, biased upward about the magnitude of η_s (i.e., $.3$) for those who endorsed 1, biased upward about 2 times the magnitude (i.e., $.6$) for those who endorsed 2, biased upward about 3 times the magnitude (i.e., $.9$) for those who endorsed 3, and so on for higher scores. Because there were 4-point scales, the bias in the estimates for the δ parameters of Scale 2 would be between 0 and $.9$.

A negative η_s would make the statements in Scale 2 more difficult, which would lower the number-correct scores of Scale 2. Because different respondents suffered from different digress of η_s , the correlation between the two scales would be underestimated when the carry-over effect was ignored by fitting the GPCM. Furthermore, ignoring η_s would affect the estimation for the α parameters of Scale 2. The variance of Scale 2 would be slightly affected sustainably when $\rho = 0$, shrunken slightly when $\rho = .4$, and shrunken substantially when $\rho = .8$. This was because the statements of Scale 2 became more difficult and the range in number-correct scores of Scale 2 would be smaller. The stronger the correlation, the larger the scale shrinkage would

be. For model identification, the variance of Scale 2 was set at 1 under the GPCM. In doing so, the shrinkage in the variance was recovered such that the α parameters of Scale 2 would be shrunk, to maintain the relationship between item responses and model parameters. In other words, ignoring η_s by fitting the GPCM would underestimate the α parameters of Scale 2 especially when ρ was large.

In Study 2, item responses were generated from the MMCEO with the constraint of Equation 17 (i.e., the carry-over effect was identical across statements, denoted as MMCEO-I), in which $\eta_{11} = -.1$, $\eta_{21} = -.3$, and $\eta_{31} = -.6$. The other settings were identical to those in Study 1. Both the MMCEO and MMCEO-L were fit to the simulated data. It was expected that the MMCEO would yield good parameter recovery, although it was unnecessarily complicated, whereas MMCEO-L would yield slightly poor parameter recovery because the linear constraint was violated to a small degree (i.e., the three values of $-.1$, $-.3$, and $-.6$ were not exactly linear). More serious violations of linearity were not manipulated (e.g., $-.1$, $-.3$, and $.5$), mainly because a linear constraint in the following empirical examples appeared appropriate. The data-generating MMCEO-I was not fit because the major purpose of Study 2 was to investigate how robust the MMCEO and MMCEO-L were. There were 100 replications.

WinBUGS was used to estimate parameters in both conditions. The following priors were set: $N(0, 10)$ for the δ and η parameters, lognormal $(0, 10)$ for the α parameters, and a uniform distribution of $(-1, 1)$ for the correlation between the two latent traits. The appendix lists the WinBUGS codes for the MMCEO. A brief simulation suggested the use of the first 5,000 iterations as burn-in, followed by an additional 5,000 iterations. Parameter estimates were sampled from the remaining 5,000 iterations per 10 values. The deviance information criterion (DIC) was used for model comparison.

Results

Study 1

The parameters in Scale 1 followed the GPCM; thus, it was expected that the three models would yield similar parameter estimates. The simulation results supported this expectation. For example, when $\rho = .4$ and $\eta_s = -.3$, the mean RMSE for the α parameters under the GPCM, MMCEO, and MMCEO-L was .058, and that for all the δ parameters was .085. In contrast, the parameter estimates for Scale 2 were very different across the three models. Tables 1 to 3 summarize the bias and RMSE values in the parameter estimates for Scale 2 of the three models, when $\rho = 0$, .4, and .8, respectively. First, consider Table 1 ($\rho = 0$). When $\eta_s = 0$ (local independence), all three models yielded very similar bias and RMSE, suggesting that fitting unnecessarily complicated models of the MMCEO and MMCEO-L did little harm, as compared with the true model of the GPCM. Furthermore, the MMCEO and MMCEO-L yielded estimates for the η parameters that were very close to their true values of zero.

When $\eta_s = -.3$ (local dependence), fitting the GPCM yielded much worse bias and RMSE for the δ parameters and the ρ parameters than fitting the MMCEO and MMCEO-L yielded. For example, the mean bias and mean RMSE for the δ parameters were .516 and .526, respectively, for the GPCM; .007 and .130, respectively, for the MMCEO; and .000 and .114, respectively, for the MMCEO-L. Here, η_s was set at $-.3$, and ignoring η_s by fitting the GPCM would overestimate the δ parameters because η_s was absorbed into the δ estimates. Although the MMCEO was unnecessarily complicated, it yielded bias and RMSE for the δ parameters that were only slightly worse than those for the data-generating MMCEO. With respect to the α parameters, all models yielded similar bias and RMSE. With respect to the η parameters, the MMCEO yielded

Table 1. Bias and RMSE in Item Parameters for the Second Scale When Two Scales Are Independent ($\rho = 0$) in Study 1.

	Nil ($\eta = 0$)						LID ($\eta = -.3$)					
	GPCM			MMCEO			GPCM			MMCEO		
	Bias	RMSE		Bias	RMSE		Bias	RMSE		Bias	RMSE	
α	Maximum	.000	.100	-.013	.104	.101	-.008	.112	-.013	.097	-.007	.097
	Minimum	-.018	.044	-.034	.047	.045	-.064	.037	-.027	.040	-.019	.038
	M	-.008	.063	-.020	.065	.063	-.039	.069	-.020	.061	-.013	.059
δ	Maximum	.020	.124	.031	.154	.149	.768	.773	.053	.193	.040	.173
	Minimum	-.018	.056	-.036	.068	.064	.196	.220	-.040	.077	-.037	.075
	M	.001	.079	-.001	.094	.087	.516	.526	.007	.130	.000	.114
η	Maximum	—	—	.018	.123	.033	—	—	.039	.203	.013	.050
	Minimum	—	—	-.014	.057	.018	—	—	-.030	.070	-.001	.029
	M	—	—	.003	.086	.027	—	—	.009	.123	.006	.037
ρ	-.001	.025	.025	.000	.025	.025	-.196	.198	.000	.029	.000	.029

Note. — constrained at zero. RMSE = root mean square error; GPCM = generalized partial credit model; MMCEO = multidimensional model for carry-over effect in OSMS; MMCEO-L = multidimensional model for carry-over effect in OSMS with linear constraint.

Table 2. Bias and RMSE in Item Parameters for the Second Scale When Two Scales Are Moderately Correlated ($\rho = .4$) in Study 1.

	Nil ($\eta = 0$)						LID ($\eta = -.3$)					
	GPCM			MMCEO			GPCM			MMCEO		
	Bias	RMSE		Bias	RMSE		Bias	RMSE		Bias	RMSE	
α	Maximum	-.007	.098	-.018	.105		-.049	.257		.003	.139	
	Minimum	-.024	.047	-.037	.049		-.225	.061		-.030	.042	
	M	-.014	.066	-.026	.070		-.103	.118		-.017	.064	
δ	Maximum	.039	.150	.058	.167		.724	.727		.048	.173	
	Minimum	-.027	.052	-.038	.055		.122	.175		-.044	.068	
	M	.000	.078	.003	.094		.442	.460		.007	.119	
η	Maximum	—	—	.017	.145		—	—		.030	.186	
	Minimum	—	—	-.026	.044		—	—		-.009	.067	
	M	—	—	-.004	.086		—	—		.009	.113	
ρ	—	-.007	.024	-.010	.026		-.206	.207		-.006	.023	

Note. — constrained at zero. RMSE = root mean square error; GPCM = generalized partial credit model; MMCEO = multidimensional model for carry-over effect in OSMS; MMCEO-L = multidimensional model for carry-over effect in OSMS with linear constraint.

Table 3. Bias and RMSE in Item Parameters for the Second Scale When Two Scales Are Highly Correlated ($\rho = .8$) in Study 1.

	Nil ($\eta = 0$)						LID ($\eta = -.3$)					
	GPCM			MMCEO			GPCM			MMCEO		
				MMCEO-L								
	Bias	RMSE		Bias	RMSE		Bias	RMSE		Bias	RMSE	
α	Maximum	—	.005	.078	.089	.007	.085	.408	.105	.025	.105	.103
	Minimum	—	-.026	.041	.047	-.037	.046	.119	.054	-.034	.054	.051
	M	—	-.015	.062	.072	-.020	.069	.243	.075	-.029	.075	.073
δ	Maximum	—	.037	.129	.288	.067	.215	.994	.273	.111	.273	.215
	Minimum	—	-.034	.055	.057	-.029	.056	.252	.063	-.036	.063	.060
	M	—	-.001	.087	.151	.010	.122	.651	.133	.020	.133	.111
η	Maximum	—	—	—	.275	.002	.057	—	.223	.030	.223	.045
	Minimum	—	—	—	.060	-.014	.032	—	.052	-.028	.052	.020
	M	—	—	—	.153	-.006	.041	—	.116	.000	.116	.031
ρ	—	-.006	.014	—	.017	-.008	.016	.131	.016	-.008	.016	.015

Note. — constrained at zero. RMSE = root mean square error; GPCM = generalized partial credit model; MMCEO = multidimensional model for carry-over effect in OSMS; MMCEO-L = multidimensional model for carry-over effect in OSMS with linear constraint.

a larger mean RMSE (.123) than the MMCEO-L did (.037). Finally, the GPCM underestimated the ρ parameter by .20.

Next, consider Tables 2 and 3, where $\rho = .4$ and $.8$, respectively. Similar conclusions as those in Table 1 can be drawn. When $\eta_s = 0$, all three models yielded similar bias and RMSE for the α and δ parameters, and the MMCEO and MMCEO-L yielded estimates for the η parameters that were very close to their true values of zero. When $\eta_s = -.3$, the GPCM overestimated the δ parameters and underestimated the α and ρ parameters; the MMCEO yielded a larger mean RMSE for the η parameters than the MMCEO-L yielded. A comparison of Tables 1 to 3 revealed that the stronger the correlation between latent traits, the worse the parameter estimates for the wrong models (i.e., MMCEO and MMCEO-L for GPCM data, and GPCM and MMCEO for MMCEO-L data). For example, when data were simulated from the GPCM, fitting the MMCEO yielded mean RMSE for the α , δ , and η parameters of .064, .094, and .086, respectively, when $\rho = 0$; .070, .094, and .086, respectively, when $\rho = .4$; and .072, .151, and .153, respectively, when $\rho = .8$. When data were simulated from the MMCEO-L, fitting the GPCM yielded mean RMSE for the α and δ parameters of .069 and .526, respectively, when $\rho = 0$; .118 and .460, respectively, when $\rho = .4$; and .243 and .651, respectively, when $\rho = .8$.

In addition to item parameter estimates, it was of great interest to compare the test reliability for the three models. There was little difference in the test-reliability estimates of Scale 1 among the three models. However, these three models yield slightly different test-reliability estimates of Scale 2. When data were simulated from the MMCEO-L, fitting the MMCEO and MMCEO-L yielded the same mean test-reliability estimate (across 100 replications) of scale 2 of .841, .831, and .898, when $\rho = 0$, $.4$, and $.8$, respectively. In contrast, these mean test reliabilities were .806, .792, and .876, respectively, when the GPCM was fit. Treating the MMCEO-L as a gold standard (because it was the data-generating model), it appeared that ignoring the carry-over effect by fitting the GPCM would yield biased estimates for the test reliability of Scale 2.

The effectiveness in model comparison for the DIC was evaluated. When data were simulated from the GPCM, the DIC of the GPCM was always smaller than that of the MMCEO and MMCEO-L across 100 replications. When data were simulated from the MMCEO-L, fitting the MMCEO-L almost always (>96%) had the smallest DIC among the three models, whereas fitting the GPCM almost always had the largest DIC. It could be concluded that the DIC was very effective in selecting correct models.

It was of interest to investigate the consequences of fitting the bi-factor model (where the LID was treated as recursive) to data that were simulated from the MMCEO (where the LID was nonrecursive). The results showed that fitting the bi-factor model yielded similar patterns as those yielded by fitting the GPCM. For example, when $\eta_s = -.3$ and $\rho = 0$, the bias in the δ parameters of Scale 2 under the bi-factor model was between .212 and .760, which was very close to the bias under the GPCM (between .196 and .768), and the variances of the 10 additional latent traits for the 10 statements were between .007 and .011, all very close to 0. Thus, it was inappropriate to treat nonrecursive LID as recursive LID.

Study 2

Table 4 summarizes the bias and RMSE values of the α and δ parameters of Scale 2 under the MMCEO and MMCEO-L. It appears that both models recovered the α parameters very well, but the MMCEO-L underestimated the δ parameters slightly by .099, .115, and .102, respectively, on average when $\rho = 0$, $.4$, and $.8$. Both models yielded almost identical test-reliability estimates. Although the MMCEO-L was not the data-generating model, it appeared robust as long as the linearity assumption was not seriously violated.

Table 4. Bias and RMSE in Item Parameters for the Second Scale for the MMCEO and MMCEO-L in Study 2.

	MMCEO						MMCEO-L					
	$\rho = 0$			$\rho = .4$			$\rho = .8$			$\rho = 0$		
	Bias	RMSE		Bias	RMSE		Bias	RMSE		Bias	RMSE	
α												
Maximum	-.010	.109		-.017	.128		-.018	.128		-.006	.106	
Minimum	-.033	.042		-.036	.041		-.044	.052		-.035	.040	
M	-.021	.072		-.026	.075		-.031	.073		-.019	.071	
δ												
Maximum	.053	.207		.045	.279		.128	.279		-.004	.316	
Minimum	-.032	.050		-.050	.060		-.042	.067		-.296	.049	
M	.002	.025		-.011	.025		-.006	.014		.001	.025	
ρ	.007	.115		.008	.116		.012	.116		-.099	.151	
										-.005	.365	
										-.283	.062	
										-.009	.024	
										-.115	.157	
										-.008	.266	
										-.223	.066	
										-.003	.013	
										-.102	.151	

Note. RMSE = root mean square error; MMCEO = multidimensional model for carry-over effect in OSMS; MMCEO-L = multidimensional model for carry-over effect in OSMS with linear constraint.

Two Empirical Examples

Example 1: School Bullying With Parallel Design

A survey on high school students' experiences in school was analyzed, in which students were asked to evaluate the frequency and severity as victims of nine bullying behaviors in school (Mok, Wang, Cheng, Leung, & Chen, 2013). In the frequency scale, there were five response categories (0 = *never*, 1 = *1 to 2 times in a half year*, 2 = *2 to 3 times in a month*, 3 = *once a week*, and 4 = *several times in a week*); in the severity scale, there were five response categories (0 = *very mild*, 1 = *mild*, 2 = *moderate*, 3 = *severe*, and 4 = *very severe*). The parallel design was adopted. A total of 5,172 students were recruited from Taiwan, Hong Kong, and Macau. The authors were interested in whether victims were inclined to exaggerate or suppress the severity of what they have often suffered from. The GPCM, MMCEO, and MMCEO-L were fit to the data.

The DIC values for the GPCM, MMCEO, and MMCEO-L were 135,068, 131,700, and 131,872, respectively, suggesting that the GPCM had the worst fit. The posterior predictive p value of Bayesian chi-square for the MMCEO and MMCEO-L was around .90, indicating a good fit. The test reliability of the frequency scale was .712, and that of the severity scale was .88, for all three models. The correlation between the two latent traits under the GPCM was .09 ($SE = .02$), whereas the correlation under the MMCEO and MMCEO-L was $-.05$ ($SE = .02$) and $-.04$ ($SE = .02$), which suggests that the direction of correlation between scales could be reversed if the carry-over effect were ignored.

Figure 1 presents the η estimates in the MMCEO. All 36 estimates were positive, and they increased as the item scores increased, suggesting a positive carry-over effect from frequency to severity. In other words, these victims were inclined to exaggerate the severity of what they had often suffered from. A close look at the nine items in Figure 1 reveals a clear linear effect of the η parameters. Although the MMCEO-L did not have a better fit than that of the MMCEO, the former model appeared practically acceptable.

From the previous simulation studies, it was found that ignoring negative carry-over effects led to overestimates of the δ parameters in the GPCM. In this example, the carry-over effects were positive, so the δ parameters in the GPCM would be underestimated. The GPCM yielded smaller estimates than the other two models yielded; the MMCEO and MMCEO-L yielded very similar estimates. Treating the estimates for the δ parameters in the MMCEO-L as a gold standard (it had a better fit), the authors found that the GPCM underestimated the δ parameters by $-.15$ to $.02$ ($M = -.07$). This underestimation indicated a positive carry-over effect.

Example 2: Verbal Aggression With Sequential Design

The verbal aggression data set (De Boeck, 2008; De Boeck & Wilson, 2004; Smits, De Boeck, & Vansteelandt, 2004) included responses of 316 persons to 12 statements (e.g., a bus fails to stop for me) on a want scale and a do scale. Each scale had three categories (0 = *no*, 1 = *to some extent*, and 2 = *to a strong extent*). The sequential design was adopted. The authors were interested in whether an endorsement on a statement along the want scale would exaggerate or suppress the endorsement on the same statement on the do scale. Due to the small sample size of 316 persons, the Rasch-type models of the partial credit model (PCM; Masters, 1982) MMCEO, and MMCEO-L were fit, which yielded a DIC value of 12115, 11778, and 11,798, respectively. Thus, the PCM had the worst fit and the MMCEO had the best fit. The posterior predictive p values of the Bayesian chi-square were .13 and .14 for the MMCEO and MMCEO-L, respectively, indicating a good fit. The correlation between the want and do latent traits was .56, indicating that they were distinct latent traits. The test reliabilities were .84, .82, and .82 for the want scale, and .86, .81, and .81 for the do scale, under the PCM, MMCEO, and MMCEO-L, respectively.

Figure 1. Estimates of the η parameters across item scores in the MMCEO of Empirical Example 1.
 Note. MMCEO = multidimensional model for carry-over effect in OSMS.

Figure 2 presents the estimates for the η parameters in the MMCEO. A total of 23 of the 24 parameter estimates were positive, suggesting positive carry-over effects. In other words, an endorsement of a high degree on the want scale tended to exaggerate the endorsement of a high degree on the do scale.

Conclusion and Discussion

In OSMS instruments, a statement is rated on multiple scales. These responses may be locally dependent because they are connected with the same statement. Because of the sequence of these responses, the resulting carry-over effect is nonrecursive: A response can be affected only by its preceding response, not by its subsequent response. Existing IRT models fail to consider the carry-over effect in OSMS instruments. In this study, the MMCEO was proposed, in which the item responses on the first scale followed standard IRT models (e.g., the GPCM), and those on the second scale have additional carry-over effect η parameters. The MMCEO can be generalized to more than two scales, and the carry-over effect across statements can be constrained when appropriate.

Figure 2. Estimates of the η parameters across item scores in the MMCEO of Empirical Example 2.
 Note. MMCEO = multidimensional model for carry-over effect in OSMS.

Two simulation studies were conducted to assess the parameter recovery of the MMCEO and the consequences of ignoring the carry-over effect on parameter estimation. Simulation results demonstrated that the parameters of the MMCEO could be recovered fairly well with WinBUGS, even when there was no carry-over effect. Moreover, fitting the unnecessarily complicated MMCEO did little harm on the parameter estimation and yielded close to zero estimates for the η parameters; ignoring the carry-over effect by fitting standard IRT models (e.g., the GPCM) yielded biased estimates for the item parameters of Scale 2, the correlation between latent traits, and the test reliability. Finally, the DIC appeared very powerful in selecting true models.

Two empirical examples were provided to demonstrate implications and applications of the MMCEO. The first example was about school-bullying behaviors, using the parallel design with frequency and severity scales. The results revealed positive carry-over effects, in which an endorsement of high frequency of a bullying behavior tended to exaggerate the next endorsement of high severity of the same bullying behavior. The latent traits of frequency and severity were nearly uncorrelated. The second example was about verbal aggression, using the sequential design. Similar results to those in the first empirical example were found. The positive carry-over effects indicated that an endorsement of high degree on the want scale tended to exaggerate the next endorsement of high degree on the do scale. It was observed that the LID in sequential design could be eliminated when there was sufficient time gap or distracting work between responding to two scales (Wang et al., 2005). The verbal aggression data, not having a time gap or distracting work between scales, revealed a carry-over effect.

The MMCEO can be further extended to describe more complicated carry-over effect. For example, different groups of respondents (e.g., males and females) may exhibit different carry-over effects when responding to OSMS instruments. Furthermore, the group membership may be latent, which calls for mixture IRT models (Rost, 1990; Rost & Langeheine, 1997). The carry-over effect η parameters can be random effects, indicating that different respondents exhibit different amounts of carry-over effect (De Boeck, 2008). Model extensions often come at the price of large amounts of data. Future studies could be conducted to develop additional general models for complicated carry-over effects and to evaluate their performance under different conditions (e.g., test length and sample size).

Appendix

WinBUGS Codes for the MMCEO of Simulation Study I

```
# N is the number of persons;
# T is the number of items in a scale;
# r is the data matrix with N rows and T*2 columns;
# alpha is the item slope;
# delta is the item threshold;
# theta is the latent trait;
# rho is the correlation between two latent traits;
# eta is the carry-over effect from the first item onto the second item
of a statement;

model {
  for (i in 1:N) {
    theta[i,1:2] ~ dmnorm(mu[1:2], I_cov[1:2, 1:2])

    # for the first scale
    for (j in 1:T) {
      Q[i,j,1] <- 1
      Q[i,j,2] <- Q[i,j,1]*exp(alpha[j]*(theta[i,1] - delta[j,1]))
      Q[i,j,3] <- Q[i,j,2]*exp(alpha[j]*(theta[i,1] - delta[j,2]))
      Q[i,j,4] <- Q[i,j,3]*exp(alpha[j]*(theta[i,1] - delta[j,3]))
      denom[i,j] <- sum(Q[i,j,])
      PP[i,j,1] <- Q[i,j,1]/denom[i,j]
      PP[i,j,2] <- Q[i,j,2]/denom[i,j]
```

```

PP[i,j,3] <- Q[i,j,3]/denom[i,j]
PP[i,j,4] <- Q[i,j,4]/denom[i,j]
r[i,j] ~ dcat(PP[i,j,])
}

# for the second scale
for (j in (T + 1):(2*T)) {
  Q[i,j,1] <- 1
  Q[i,j,2] <- Q[i,j,1]*exp(alpha[j]*(theta[i,2] - delta[j,1] +
eta[j-T,r[i,j-T]]))
  Q[i,j,3] <- Q[i,j,2]*exp(alpha[j]*(theta[i,2] - delta[j,2] +
eta[j-T,r[i,j-T]]))
  Q[i,j,4] <- Q[i,j,3]*exp(alpha[j]*(theta[i,2] - delta[j,3] +
eta[j-T,r[i,j-T]]))
  denom[i,j] <- sum(Q[i,j,])
  PP[i,j,1] <- Q[i,j,1]/denom[i,j]
  PP[i,j,2] <- Q[i,j,2]/denom[i,j]
  PP[i,j,3] <- Q[i,j,3]/denom[i,j]
  PP[i,j,4] <- Q[i,j,4]/denom[i,j]
  r[i,j] ~ dcat(PP[i,j,])
}
}

# Priors
mu[1] <- 0
mu[2] <- 0

rho ~ dunif(-1, 1)
covm[1,1] <- 1
covm[1,2] <- rho
covm[2,1] <- rho
covm[2,2] <- 1
I_cov[1:2, 1:2] <- inverse(covm[1:2, 1:2])

for (j in 1:(2*T)) {
  alpha[j] ~ dlnorm(0, 0.1)
  for (k in 1:3) {
    delta[j,1] ~ dnorm(0, 0.1)
    delta[j,2] ~ dnorm(0, 0.1)
    delta[j,3] ~ dnorm(0, 0.1)
  }
}

for (j in 1:T) {
  eta[j,1] <- 0
  eta[j,2] ~ dnorm(0, 0.1)
  eta[j,3] ~ dnorm(0, 0.1)
  eta[j,4] ~ dnorm(0, 0.1)
}
}

```

Note. MMCEO = multidimensional model for carry-over effect in OSMS.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Andrich, D., Humphry, S. M., & Marais, I. (2012). Quantifying local, response dependence between two polytomous items using the Rasch model. *Applied Psychological Measurement, 36*, 309-324. doi:10.1177/0146621612441858
- Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement, 34*, 181-192. doi:10.1177/0146621609360202
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51. doi:10.1007/bf02291411
- Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*, 581-612. doi:10.1007/s11336-010-9178-0
- Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289. doi:10.2307/1165285
- Cummings, J. L., Mega, M., Gray, K., Rosenberg-Thompson, S., Carusi, D. A., & Gornbein, J. (1994). The Neuropsychiatric Inventory: Comprehensive assessment of psychopathology in dementia. *Neurology, 44*, 2308-2314.
- De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533-559. doi:10.1007/S11336-008-9092-X
- De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.
- Hoskens, M., & Boeck, P. D. (1995). Componential IRT models for polytomous items. *Journal of Educational Measurement, 32*, 364-384. doi:10.2307/1435218
- Hoskens, M., & De Boeck, P. (2001). Multidimensional componential item response theory models for polytomous items. *Applied Psychological Measurement, 25*, 19-37. doi:10.1177/01466216010251002
- Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*, 3-21. doi:10.1177/0146621605275414
- Liu, Y., & Verkuilen, J. (2013). Item response modeling of presence-severity items: Application to measurement of patient-reported outcomes. *Applied Psychological Measurement, 37*, 58-75. doi:10.1177/0146621612455091
- Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*, 200-215.
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174. doi:10.1007/BF02296272
- Mok, M. M. C., Wang, W.-C., Cheng, Y.-Y., Leung, S.-O., & Chen, L.-M. (2013). Prevalence and behavioral ranking of bullying and victimization among secondary students in Hong Kong, Taiwan, and Macao. *The Asia-Pacific Education Researcher, 23*, 757-767. doi:10.1007/s40299-013-0151-4
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176. doi:10.1177/014662169201600206
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282. doi:10.1177/014662169001400305

- Rost, J., & Langeheine, R. (1997). *Applications of latent trait and latent class models in the social sciences*. Muster, Germany: Waxman.
- SAS Institute. (1999). The NLMIXED Procedure [Computer program]. Cary, NC: Author.
- Smits, D. J. M., De Boeck, P., & Vansteelandt, K. (2004). The inhibition of verbally aggressive behavior. *European Journal of Personality*, 18, 537-555. doi:10.1002/per.529
- Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2007). *WinBUGS version 1.4.3*. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health. Retrieved from <http://www.mrc-bsu.cam.ac.uk/bugs>
- Tuerlinckx, F., & De Boeck, P. (1999). Distinguishing constant and dimension-dependent interaction: A simulation study. *Applied Psychological Measurement*, 23, 299-307. doi:10.1177/01466219922031419
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.
- Wang, W.-C., Cheng, Y.-Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement*, 65, 5-27. doi: 10.1177/0013164404268676
- Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, 29, 126-149. doi:10.1177/0146621604271053
- Wilson, M., & Adams, R. (1995). Rasch models for item bundles. *Psychometrika*, 60, 181-198. doi: 10.1007/BF02301412
- Yousfi, S., & Böhme, H. F. (2012). Principles and procedures of considering item sequence effects in the development of calibrated item pools: Conceptual analysis and empirical illustration. *Psychological Test and Assessment Modeling*, 54, 366-396.