Bulut *et al. Large-scale Assess Educ (2017) 5:8*
DOI 10.1186/s40536-017-0042-x

Large-scale Assessments in Education

METHODOLOGY Open Access

A structural equation modeling approach for examining position effects in large-scale assessments

Okan Bulut*, Qi Quo and Mark J. Gierl

*Correspondence: bulut@ualberta.ca
Centre for Research in Applied Measurement and Evaluation, University of Alberta, 6-110 Education Centre North, 11210 87 Ave NW, Edmonton, AB T6G 2G5, Canada

Abstract

Position effects may occur in both paper–pencil tests and computerized assessments when examinees respond to the same test items located in different positions on the test. To examine position effects in large-scale assessments, previous studies often used multilevel item response models within the generalized linear mixed modeling framework. Using the equivalence of the item response theory and binary factor analysis frameworks when modeling dichotomous item responses, this study introduces a structural equation modeling (SEM) approach that is capable of estimating various types of position effects. Using real data from a large-scale reading assessment, the SEM approach is demonstrated for investigating form, passage position, and item position effects for reading items. The results from a simulation study are also presented to evaluate the accuracy of the SEM approach in detecting item position effects. The implications of using the SEM approach are discussed in the context of large-scale assessments.

Keywords: Position effect, Large-scale assessment, Structural equation modeling, Item response theory

Background

Large-scale assessments in education are typically administered using multiple test forms or booklets in which the same items are presented in different positions or locations within the forms. The main purpose of this practice is to improve test security by reducing the possibility of cheating among test takers (Debeer and Janssen 2013). This practice also helps test developers administer a greater number of field-test items embedded within multiple test forms. Although this is an effective practice for ensuring the integrity of the assessment, it may result in context effects—such as an item position effect—that can unwittingly influence the estimation of item parameters and the latent trait (Bulut 2015; Hohensinn et al. 2011). For example, test takers may experience either increasing item difficulty at the end of the test due to fatigue or decreasing item difficulty due to test-wiseness as they become more familiar with the content (Hohensinn et al. 2008).

It is often assumed that position effects are the same for all test takers or for all items and thus do not have a substantial impact on item difficulty or test scores (Hahne 2008).

However, this assumption may not be accurate in operational testing applications, such as statewide testing programs. For example, in a reading assessment where items are often connected to a reading passage, it is challenging to maintain the original positions of the items from field test to final form to eliminate the possibility of any position effects (Meyers et al. 2009). Similarly, the positions of items cannot be controlled in computerized adaptive tests in which item positions typically differ significantly from one examinee to another (Davey and Lee 2011; Kolen and Brennan 2014; Meyers et al. 2009). Therefore, non-negligible item position effects may become a source of error in the estimation of item parameters and the latent trait (Bulut et al. 2016; Hahne 2008).

Several methodological studies have focused on item position effects in national and international large-scale assessments, such as the Trends in International Mathematics and Science Study (Martin et al. 2004), a mathematics assessment of the Germany Educational Standards (Robitzsch 2009), a mathematical competence test of the Austrian Educational Standards (Hohensinn et al. 2011), the Graduate Record Examination (Albano 2013), the Program for International Student Assessment (Debeer and Janssen 2013; Hartig and Buchholz 2012), and a German nationwide large-scale assessment in mathematics and science (Weirich et al. 2016). In these studies, item position effects were estimated as either fixed or random effects using multilevel item response theory (IRT) models within the generalized linear mixed modeling framework. This approach is also known as explanatory item response modeling (De Boeck and Wilson 2004), which is equivalent to Rasch modeling (Rasch 1960) with explanatory variables.

An alternative way of estimating IRT models is with a binary factor analysis (FA) model in which the item parameters and latent trait can be estimated using tetrachoric correlations among dichotomous item responses (Kamata and Bauer 2008; McDonald 1999; Takane and de Leeuw 1987). In addition to estimating the item parameters and latent trait, the binary FA model can be expanded to a structural equation model (SEM) in which item difficulty parameters can be predicted by other manifest (i.e., observed) variables (e.g., item positions, passage or testlet positions, and indicators of test forms). The purpose of this study is to (1) introduce the SEM approach for detecting position effects in large-scale assessments; (2) demonstrate the methodological adequacy of the SEM approach to model different types of position effects using an empirical study; and (3) examine the accuracy of the SEM approach in detecting position effects using a simulation study.

Theoretical framework

IRT and factor analysis

The FA model is a regression model in which the independent variables are latent variables and the dependent variables are manifest variables (Ferrando and Lorenzo-Seva 2013). Previous studies on the relationship between the FA and IRT frameworks showed that the unidimensional two-parameter normal ogive IRT model is equivalent to a one-factor FA model when binary manifest variables are predicted by a continuous latent variable (e.g., Brown 2006; Ferrando and Lorenzo-Seva 2005; Glöckner-Rist and Hoijtink 2003; MacIntosh and Hashim 2003; Takane and de Leeuw 1987). The two-parameter normal ogive IRT model can be written as

$$P(y_i = 1|\theta, a_i, b_i) = \Phi(a_i\theta - b_i), \tag{1}$$

where y_i is the dichotomous item responses for item i, θ is the latent trait (i.e., ability), a_i is the item discrimination parameter, b_i is the item difficulty parameter, and Φ is the standard normal cumulative distribution function. Using the parameterization described in Asparouhov and Muthén (2016), the one-factor FA model for binary manifest variables that corresponds to the two-parameter IRT model in Eq. 1 can be written as follows:

$$y_i^* = \nu_i + \lambda_i \eta + \varepsilon_i, \tag{2}$$

where y_i^* is the continuous latent response underlying the observed item response for item i, ν_i is the intercept term for item i, which is typically assumed to be zero, λ_i is the factor loading for item i, which is analogous to item discrimination in traditional IRT models, η is the latent trait, which is often assumed to follow a normal distribution as $\eta \sim N(0, 1)$[1], and ε_i is the residual term[2] for item i (for compactness in the notation, no person subscript is included). In case of the Rasch model (Rasch 1960) or the one-parameter IRT model, factor loadings in the one-factor FA model are fixed to "1" or constrained to be equal across all items.

In the one-factor FA model, there is also a threshold parameter (τ_i) for each dichotomous item, which corresponds to item difficulty (b_i in Eq. 1) in traditional IRT models. Based on the threshold parameter, an observed item response y_i becomes

$$y_i = \begin{cases} 1, & \text{if} \quad y_i^* \geq \tau_i \\ 0, & \text{if} \quad y_i^* < \tau_i. \end{cases}$$

Although factor loadings and intercepts in the one-factor FA model are analogous to item discrimination and item difficulty in IRT, these factor analytic terms can be formally transformed into item parameters in the traditional IRT scale (Asparouhov and Muthén 2016; Muthén and Asparouhov 2002). Assuming that the latent trait is normally distributed as $\eta \sim N(\alpha, \psi)$ and $\eta = \alpha + \sqrt{\psi}\theta$ where θ is the IRT-based latent trait with mean 0 and standard deviation 1, the item discrimination parameter can be computed as $a_i = \lambda_i / \sqrt{\psi}$, and the item difficulty parameter can be computed as $b_i = (\tau_i - \lambda_i \alpha) / \left(\lambda_i \sqrt{\psi}\right)$ [see Brown (2006) and Takane and de Leeuw (1987) for a review of similar transformation procedures].

Modeling position effects

As Brown (2006) noted, the use of the FA model provides greater analytic flexibility than the IRT framework because traditional IRT models can be embedded within a larger model that includes additional variables to explain the item parameters as well as the latent trait (e.g., Ferrando et al. 2013; Glöckner-Rist and Hoijtink 2003; Lu et al. 2005). Using the structural equation modeling (SEM) framework, an IRT model can be defined as a *measurement* model in which there is a latent trait (e.g., reading ability) underlying a set of manifest variables (e.g., dichotomous items in a reading assessment). In the *structural* part of the SEM model, the causal and correlational relations among the latent trait, the manifest variables, and other latent or manifest variables (e.g., gender and attitudes toward reading) can therefore be tested.

[1] The mean and variance are fixed to 0 and 1, respectively, to identify the scale of the latent variable.

[2] The residuals are typically assumed to have a normal distribution (Kamata and Bauer 2008).

A well-known example of this modeling framework is the Multiple Indicators Multiple Cause (MIMIC) model for testing uniform and nonuniform differential item functioning in dichotomous and polytomous items (e.g., Finch 2005; Lee et al. 2016; Woods and Grimm 2011). In the MIMIC model, a categorical grouping variable (e.g., gender) is used as an explanatory variable to explain the relationship between the probability of responding to an item correctly and the grouping variable, after controlling for the latent trait. Based on the results of this analysis, one can conclude that the items become significantly less or more difficult depending on which group the examinee belongs to.

In the current study, instead of a categorical grouping variable, *item position* will be used as a continuous, explanatory variable to predict whether there is any relationship between item difficulty and the varying positions of the items on the test. Using the FA model in Eq. 2, the SEM model for examining linear item position effects can be written as follows:

$$y_i^* = \lambda_i \eta + \beta_i p_i + \varepsilon_i, \tag{3}$$

where p_i is a manifest variable that shows the positions of item i across all examinees who responded to item i (e.g., 1, 15, 25, and 50 if item i is administered at the 1st, 15th, 25th, or 50th positions), β_i is the linear position effect based on one unit change in the position of item i, and the remaining terms are the same as those from Eq. 2. In Eq. 3, if $\beta_i = 0$, it can be concluded that the difficulty of item i does not depend on its position on the test. If, however, $\beta_i \neq 0$, then the difficulty of item i either linearly increases or decreases as a function of its position on the test (i.e., p_i).

It is also possible to extend the SEM model in Eq. 3 to examine the interaction between the latent trait and item position, if the relation between the difficulty of an item and its position on the test is assumed to be moderated by the latent trait. To estimate the interaction effect, an interaction term as a product of item position and the latent trait needs to be included as an additional predictor in the model (Moosbrugger et al. 2009). The SEM model for examining linear position effects and interaction effects can therefore be written as follows:

$$y_i^* = \lambda_i \eta + \beta_i p_i + \omega_i \eta p_i + \varepsilon_i, \tag{4}$$

where p_i is the positions of item i across all examinees who responded to item i, β_i is the linear position effect, ω_i is the interaction effect for item i, and the other terms are the same as those in Eq. 3. If, $\omega_i \neq 0$, then the difficulty of item i increases or decreases as a result of the interaction between the latent trait and item positions (i.e., ηp_i in Eq. 4). There are several methods used for estimating the interaction between the latent trait and manifest variables (e.g., item and testlet positions), such as the Latent Moderated Structural Equations (LMS) approach (Klein and Moosbrugger 2000), the Quasi-Maximum Likelihood (QML) approach (Klein and Muthén 2007), the constrained approach (Jöreskog and Yang 1996), and the unconstrained approach (Marsh et al. 2004). The examples of interaction effects in SEM with binary manifest variables can be found in Woods and Grimm (2011) and Lee et al. (2016).

Figure 1 shows a path diagram of an SEM model based on k dichotomous items. In the model, item 1 is tested for linear position effects and interaction effects. Although Fig. 1 only shows the estimation of item position effect for item 1, all items on the test (or a

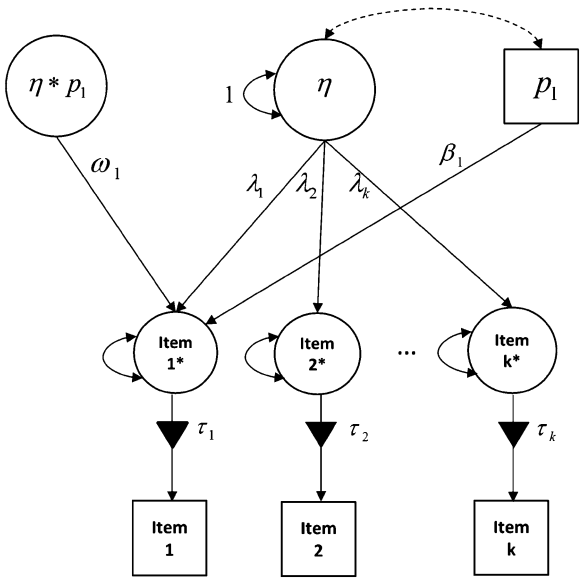

Fig. 1 The structural equation model for examining item position effects (*Note* η = Latent trait; λ_1 to λ_k = factor loadings; τ_1 to τ_k = item thresholds; p_1 = positions of item 1 on the test; β_1 = linear item position effect for item 1; ω_1 = interaction term between latent trait and positions of item 1)

group of items) could be evaluated together for item position effects in the same model. It should be noted that the correlations between item positions and the latent trait (i.e., a double-headed arrow as in Fig. 1) are optional. If item positions are randomly determined for each examinee (i.e., independent of examinees' latent trait levels), then the correlation between item positions and the latent trait can be excluded from the model for the sake of model parsimony. However, the ordering of items can sometimes depend on examinees' latent trait level. For example, in a computerized adaptive test, two examinees can receive the same item at different positions because of how they responded to earlier items on the test (e.g., van der Linden et al. 2007). Furthermore, in a fixed-form and non-adaptive test, examinees can still start and exit the test with different items, depending on test administration rules related to examinees' response patterns (e.g., Bulut et al. 2016; Li et al. 2012). When examining item position effects in such assessments, a correlational link between item positions and the latent trait should be included in the model to account for this dependency.

Model constraints and assumptions

The proposed SEM approach requires the use of constraints to ensure model identification and accurate estimation of model parameters. Using the marginal parameterization with a standardized factor (Asparouhov and Muthén 2016; Millsap and Yun-Tein 2004; Muthén and Asparouhov 2002), the variance of y_i^* is constrained to be 1 for all items; the mean and the variance of the latent trait (η) are constrained to be 0 and 1, respectively [see Kamata and Bauer (2008) for other parameterizations in the binary FA model]. Furthermore, additional constraints might be necessary depending on which the IRT model is used. For example, if the Rasch model is the underlying IRT model for the data, all factor loadings in the model must be constrained to be 1. The general assumptions of

IRT must also hold in order to estimate the IRT model with item position effects within the SEM framework. These assumptions include a monotonic relationship between the probability of responding to an item correctly and the latent trait, the unidimensionality of the latent trait, local independence of items, and invariance in the item parameters and the latent trait across different subgroups in a population.

Model estimation

The proposed SEM models for examining item position effects can be estimated using commercial software programs for the SEM analysis, such as Mplus (Muthén and Muthén 1998–2015), LISREL (Jöreskog and Sörbom 2015), AMOS (Arbuckle 2011), and EQS (Bentler and Wu 2002), or non-commercial software programs, such as the *sem* (Fox et al. 2016), *lavaan* (Rosseel 2012), *OpenMx* (Pritikin et al. 2015), and *nlsem* (Umbach et al. 2017) packages in R (R Core Team 2016). It should be noted that these software programs differ with regard to their algorithms for estimating SEM models with binary variables, model estimators (e.g., ULS, MLR, and WLSMV), the capability to estimate interaction effects, and methods for handling missing data. Therefore, when choosing the most suitable program, researchers should consider the research questions that they aim to address, the statistical requirements of their hypothesized SEM model(s), and data characteristics (e.g., the number of items, amount of missing data).

As noted earlier, item position effects can be examined one item at a time using Eqs. 3 and 4. However, it is more convenient to estimate the position effects for all items within the same SEM model and then create a simplified model by removing non-significant effects from the model. Since the simplified model would be nested within the original model that includes the position effects for all items, the two models can be compared using a Chi square (χ^2) difference test. The procedure for the χ^2 difference test varies depending on which model estimator (e.g., the robust maximum likelihood estimator 'MLR' or the weighted least square with mean- and variance-adjusted Chi square 'WLSMV' in Mplus) is used for estimating the SEM model. For example, the χ^2 difference test using the corrected loglikelihood values or the Satorra–Bentler χ^2 statistic (Satorra and Bentler 2001) are widely used when the model estimator is MLR. The reader is referred to Brown (2006, p. 385), Asparouhov and Muthén (2006), and the Mplus website (www.statmodel.com/chidiff.shtml) for detailed descriptions of the χ^2 difference testing in SEM. When the SEM models are not nested, the model selection or comparison can be done on the basis of information-based criteria that assess relative model fit, such as Akaike information criterion (AIC; Akaike 1974) and the Bayesian information criterion (BIC; Schwarz 1978).

Comparison with other approaches

To date, three different methodological approaches for investigating item position effects have been described: (1) logistic regression models (e.g., Davey and Lee 2011; Pomplun and Ritchie 2004; Qian 2014); (2) multilevel models based on the generalized linear mixed modeling (GLMM) framework (e.g., Albano 2013; Alexandrowicz and Matschinger 2008; Debeer and Janssen 2013; Hartig and Buchholz 2012; Li et al. 2012; Weirich et al. 2014); and (3) test equating methods (e.g., Kingston and Dorans 1984; Kolen and Harris 1990; Moses et al. 2007; Pommerich and Harris 2003; Meyers et al. 2009; Store

2013). Although there are some empirical studies that used the factor analytic methods for modeling position effects (e.g., Bulut et al. 2016; Schweizer 2012; Schweizer et al. 2009), the current study represents the first study that utilized the SEM framework as a methodological approach for examining item position effects.

The proposed SEM approach has four noteworthy advantages over the other methods mentioned above when it comes to modeling item position effects. First, the proposed approach overcomes the limitation of examining position effects only for dichotomous items, which is the case with the approaches based on the GLMM framework (e.g., Hartig and Buchholz 2012). Using the SEM framework, assessments that consist of polytomously scored items can also be examined for item position effects. Second, the proposed approach is applicable to assessments in which item parameters are obtained using the two-parameter IRT model. Because the one-factor FA model is analogous to the two-parameter IRT model, it is possible to estimate item position effects when both item difficulty and item discrimination parameters are present in the model. Third, the proposed approach can be used with multidimensional test structures in which there are multiple latent traits underlying the data. Fourth, once significant item position effects are detected, other manifest and/or latent variables (e.g., gender, test motivation, and test anxiety) can be incorporated into the SEM model to explain the underlying reasons of the found effects. For example, Weirich et al. (2016) recently found in an empirical study that item position effects in a large-scale assessment were affected by the examinees' test-taking efforts. Response time effort (Wise and Kong 2005) and disability status (Abedi et al. 2007; Bulut et al. 2016) are other important factors highlighted in the literature.

Next, the results from an empirical study first are presented to demonstrate the use of the proposed SEM approach for investigating three types of position effects that are likely to occur in large-scale assessments: test form (or booklet) effect, passage (or testlet) position effect, and item position effect. Real data from a large-scale reading assessment are used for the empirical study. The interpretation of the estimated position effects and model comparisons are explained. Then, the results from a Monte Carlo simulation study are presented to investigate the extent to which the proposed SEM approach can detect item position effects accurately. For both the empirical and Monte Carlo studies, Mplus (Muthén and Muthén 1998–2015) is used because of its flexibility to formulate and evaluate IRT models within a broader SEM framework and its extensive Monte Carlo simulation capabilities (e.g., Glöckner-Rist and Hoijtink 2003; Lu et al. 2005).

Empirical study

Data

Position effects were evaluated using data from a large-scale statewide reading assessment administered annually to all students in elementary, middle, and high schools. The assessment was delivered as either a computer-based test or a paper-and-pencil test, depending on the availability of computers in the schools. The paper-and-pencil version was administered to the students using a single test form that consisted of the same items in the same positions. Unlike the paper–pencil version, the computer-based version was administered to the students by randomizing the positions of the items across

the students. The sample used in this study consisted of 11,734 third-grade students who completed the reading test of the statewide assessment on a computer.

Instrument

The reading test consisted of 45 multiple-choice items related to seven reading passages. The positions of reading passages and items varied across students, while the positions of the items within each passage remained unchanged. This process resulted in four different patterns of item positions on the test, which were referred to as *test forms* in this study. Table 1 shows the demographic characteristics of the students across the four forms. The forms were similar in terms of sample sizes and students' demographic characteristics.

Model formulations

Four SEM models were specified and analyzed. The first model (M_1) was the baseline model because it did not include any predictors for item position effects. M_1 was equivalent to the Rasch model because the item parameters in the selected reading assessment were originally calibrated using the Rasch model. The remaining three models (i.e., M_2 to M_4) were designed to evaluate form effects, passage position effects, and item position effects, respectively. The proposed models are shown in Fig. 2. In each model, a unidimensional latent variable (i.e., reading ability) predicted a set of manifest variables (i.e., reading items). For models M_2–M_4, there were also additional predictors for the form, passage position, and item position effects.

Rasch model (M_1)

The first model was the Rasch model. It did not include any predictors for examining positions effects. Under the SEM framework, M_1 is equivalent to a one-factor model with all factor loadings fixed to 1 (see Fig. 1a). The latent variable defines the reading ability and the intercepts of the items are equivalent to item difficulties.

Form effects model (M_2)

This model examined the form effects using indicators of test forms as categorical predictors. The SEM formulation of M_2 can be written as follows:

Table 1 Demographic summary of the students across test forms

Variable	Form 1		Form 2		Form 3		Form 4	
	N	%	*N*	%	*N*	%	*N*	%
Gender								
Female	1460	49.1	1471	49.2	1477	50	1419	49
Male	1515	50.9	1517	50.8	1479	50	1478	51
Ethnicity								
American Indian	50	1.7	63	2.1	34	1.2	57	2
Asian	65	2.2	61	2	62	2.1	65	2.2
Black	272	9.1	259	8.7	236	8	249	8.6
Hispanic	373	12.5	465	15.6	428	14.5	398	13.7
White	2215	74.5	2140	71.6	2196	74.3	2128	73.5

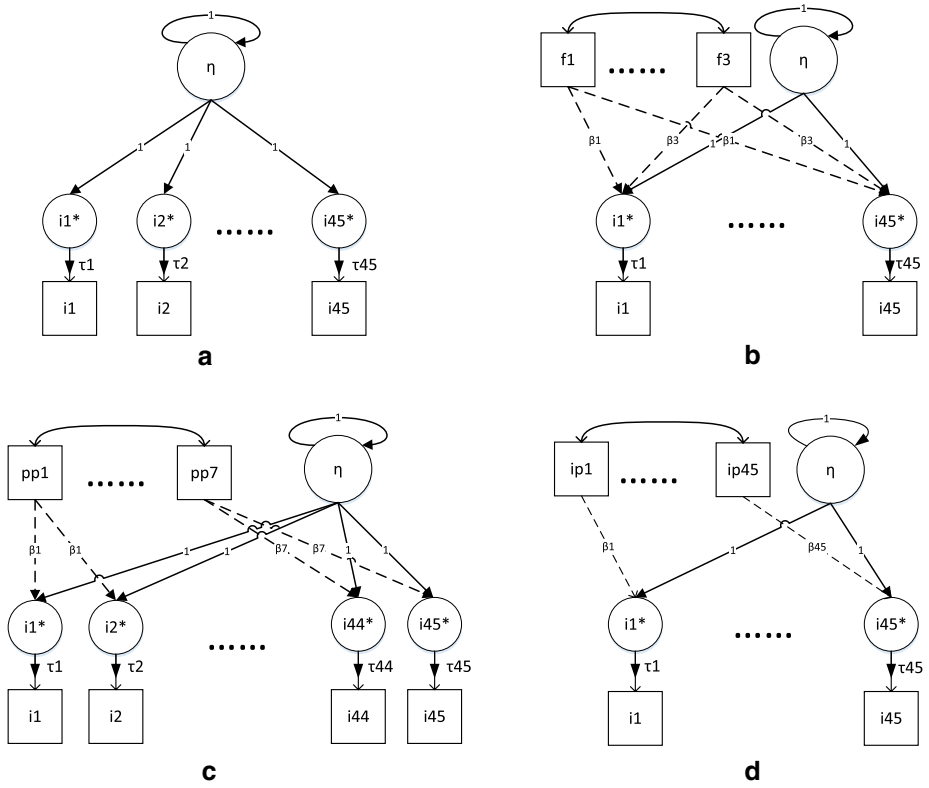

Fig. 2 Structural equation models to estimate **a** the Rasch model, **b** form effects, **c** passage position effects, and **d** item position effects (*Note*: In all figures, *i* stands for item; in **b**, *f* stands for manifest variables for form effects; in **c**, *pp* stands for manifest variables for passage position effects; and in **d**, *ip* stands for manifest variables for item position effects)

$$y_i^* = \lambda_i \eta + \beta_1 \text{form}_1 + \beta_2 \text{form}_2 + \beta_3 \text{form}_3 + \varepsilon_i, \tag{5}$$

where form_1, form_2, and form_3 are the dummy codes for the three test forms (using the fourth test form as the reference form), β_1, β_2, and β_3 are the regression coefficients that represent the differences between the overall difficulty of the three test forms and the reference test form. Note that β_1, β_2, and β_3 do not have the subscript for items, indicating that the form effects were constrained to be the same across all items. Under the SEM framework, M_2 is equivalent to a one-factor model with all factor loadings fixed to 1 and with form_1 to form_3 as additional predictors (see Fig. 1b).

Passage position effects model (M_3)

This model examined passage position effects by incorporating passage positions into the SEM model as predictors. Given item i belongs to passage h, the SEM formulation of M_3 can be shown as

$$y_i^* = \lambda_i \eta + \beta_h * \text{passage position}_h + \varepsilon_i, \tag{6}$$

where $\text{passage position}_h$ is the position of passage h across all examinees and β_h is the regression coefficient that represents the impact of one unit change in the position of passage h on the difficulty of items linked to passage h. Note that the passage position effect was constrained to be the same for all items linked to passage h. Under the SEM

Table 2 The layout of the data structure for the SEM analysis

Examinees	Item responses			Forms				Passage positions			Item positions		
	i1	…	i45	f_1	f_2	f_3	f_4	pp_1	…	pp_7	ip_1	…	ip_{45}
1				1	0	0	0	1	…	7	46	…	4
2				0	1	0	0	2	…	4	45	…	3
3				0	0	1	0	1	…	6	52	…	5
4				0	0	0	1	3	…	4	52	…	6

i1 to i45: dichotomous item responses; f1 to f4: dummy variables for the test forms; pp_1 to pp_7: passage positions; ip_1 to ip_{45}: item positions. Only one examinee from each test form was used to demonstrate the data layout

framework, M_3 is equivalent to a one-factor model with all factor loadings fixed to 1 and with passage position$_h$ as an additional predictor (see Fig. 1c).

Item position effects model (M_4)

The last model examined individual item position effects by incorporating item positions into the model as predictors. The SEM formulation of M_4 is given as:

$$y_i^* = \lambda_i \eta + \beta_i * \text{item position}_i + \varepsilon_i, \tag{7}$$

where item position$_i$ is the position of item i across all examinees, and β_i is the regression coefficient that represents the impact of one unit change in the position of item i on the difficulty of item i. Unlike the first three models summarized above, this model allows the effect of item position on each item to vary across items. Under the SEM framework, M_4 is equivalent to a one-factor model with all factor loadings fixed to 1 and with item position$_i$ as an additional predictor (see Fig. 1d).

Model estimation

Table 2 demonstrates the layout of the data structure for estimating the four SEM models (M_1 to M_4) summarized above. The SEM models were estimated using the maximum likelihood estimation with robust standard errors 'MLR' in Mplus 7 (Muthén and Muthén 1998–2015). The MLR estimator was chosen over the maximum likelihood estimator because it adjusts the standard errors of the parameter estimates when the data do not meet the multivariate normality assumption, which is often the case with binary and ordinal manifest variables (Li 2016). Mplus also provides alternative model estimators—such as WLSMV[3] and MLM—for the estimation of FA and SEM models with binary manifest variables [see Brown (2006), chapter 9, for a discussion of the estimators for categorical data]. The Mplus codes for estimating the passage and item position effects are provided in "Appendices 1 and 2".

All of the SEM models converged in less than two minutes. The estimated regression coefficients for the form, passage, and item position effects were evaluated. Then, the baseline model (i.e., M_1) was compared to the other SEM models (M_2 to M_4) using a Chi square difference test based on loglikelihood values and scaling correction factors obtained from the SEM models. The scaling correction term ($c_{correction}$) and the adjusted Chi square value ($\chi^2_{adjusted}$) for the difference test can be computed as follows:

[3] When the WLSMV estimator was used, the SEM models in the empirical study did not converge due to high correlations among passage and item position variables.

$$c_{correction} = \frac{(df_0 * c_0 - df_1 * c_1)}{(df_0 - df_1)}, \text{ and} \tag{8}$$

$$\chi^2_{adjusted} = \frac{-2(L_0 - L_1)}{c_{correction}}, \quad \text{with } df_{adjusted} = \left(df_0 - df_1\right), \tag{9}$$

where df is the number of parameters in the model, c is the scaling correction factor, L is the loglikelihood value from the estimated model, and subscripts 0 and 1 represent the baseline (i.e., M_1) and the comparison models (either M_2, M_3, or M_4), respectively. Finally, the relative fit of the non-nested SEM models (M_2, M_3, and M_4) was compared using the AIC and BIC model-fit indices.

Results

Rasch model

The Rasch model (M_1) was the baseline model without any predictors for form, passage position, or item position effects. Table 3 shows the descriptive statistics of the latent trait estimates obtained from the Rasch model across the four test forms. The distributions of the latent trait from the four test forms were similar when the form effects, passage position effects, and item position effects were ignored.

Form effects

M_2 included the indicators for the test forms in the SEM model to examine the overall differences in the item difficulty levels across the four forms. Using one form as the reference form, the estimated regression coefficients indicated the overall difficulty differences between the reference form and the other three forms. The estimated regression coefficients for the form effects ranged from −0.032 to 0.032. None of the form effects was statistically significant at the alpha level of $\alpha = .05$, suggesting that the forms did not differ in terms of their overall difficulty.

Passage position effects

Table 4 shows the estimated passage positions effects from M_3. The results indicated that passages 1, 4, and 5 showed significant passage position effects (−.029, −.044, and −.08, respectively). For example, if the position of passage 5 is increased by 1, then difficulties of the items linked to passage 5 would increase by .08 logit. Using the estimated passage position effect, the adjusted difficulty of each item linked to passage 5 can be computed. For example, the first item associated with passage 5 has a difficulty of −2.228. For the

Table 3 Descriptive statistics of latent trait estimates from the Rasch model across four test forms

Form	*N*	*M*	*SD*	Min	Max
1	2975	−.01	1.02	−3.17	2.06
2	2901	.01	1.04	−2.75	2.06
3	2897	.00	1.01	−2.75	2.06
4	2961	−.01	1.04	−3.02	2.06

Table 4 Summary of estimated passage position effects

Passage	Number of items	Position effect
1	6	−.029 (.014)*
2	6	.000 (.021)
3	8	.023 (.013)
4	6	−.044 (.015)**
5	6	−.080 (.014)***
6	6	.004 (.014)
7	7	−.008 (.008)

The order of passages in Form 1 was used as the reference. Standard errors of the estimated effects are shown in the parentheses

* p < .05, ** p < .01, *** p < .001

students who received passage 5 in the first position on the test, the estimated difficulty of the item becomes $-2.228 + (1 * .08) = -2.148$. If the same passage is administered in the fifth position, the difficulty of the same item becomes $-2.228 + (5 * .08) = -1.828$. Using the estimated position effects in Table 4, the impact of changes in passage positions on item difficulty can be calculated for other items and passages in a similar fashion.

Item position effects

Table 5 shows the estimated item positions effects from M_4. The results indicated that 23 out of the 45 items showed significant item position effects. The regression coefficients for item position effects ranged from −.009 to −.037. All of the significant position effects were negative. This finding suggests that as the positions of the items increase (e.g., the item moves from the 1st position to the 15th position), the difficulties of the items also increase and thus the probability of correctly answering the items decrease. The largest item position effect was −.037 for item 26, which had an estimated difficulty

Table 5 Summary of estimated item position effects

Item	Position effect	Item	Position effect	Item	Position effect
1	−.008 (.005)	16	−.009 (.003)**	31	.006 (.004)
2	−.015 (.005)**	17	−.012 (.003)***	32	.006 (.004)
3	−.025 (.005)***	18	−.003 (.003)	33	−.007 (.005)
4	.000 (.004)	19	−.011 (.003)**	34	.001 (.004)
5	−.014 (.005)**	20	−.015 (.003)***	35	−.007 (.004)
6	.009 (.005)	21	−.005 (.005)	36	−.005 (.004)
7	−.004 (.005)	22	.007 (.005)	37	−.010 (.004)*
8	−.023 (.005)***	23	−.014 (.005)**	38	−.008 (.004)
9	.003 (.006)	24	−.015 (.005)**	39	−.011 (.004)**
10	−.017 (.006)**	25	.002 (.006)	40	−.014 (.006)*
11	−.023 (.006)***	26	−.037 (.007)***	41	−.015 (.004)***
12	−.030 (.007)***	27	−.020 (.005)***	42	−.005 (.004)
13	−.014 (.006)*	28	−.004 (.005)	43	−.023 (.005)***
14	−.013 (.006)*	29	−.002 (.004)	44	−.006 (.003)
15	−.008 (.003)*	30	−.005 (.004)	45	−.004 (.004)

* p < .05; ** p < .01; *** p < .001

of -3.244. If item 26 was administered in the first position on the test, then the estimated difficulty of the item would become $-3.244 + (1 * .037) = -3.207$. If, however, the same item was administered at the 45th position (i.e., the last question) on the test, then the estimated difficulty of the item would become $-3.244 + (45 * .037) = -1.579$. This finding suggests that changes in the position of this item could substantially increase or decrease the probability of answering the item correctly. Adjusted difficulties of other items with regard to their positions on the test can be calculated in the same way.

Model comparison

Table 6 shows model fit information from the SEM models. In addition to the four models described earlier, there was a variant of M_4 that only included the significant item position effects in M_4. Using the Chi square difference test described in Eqs. 8 and 9, the SEM models with either passage position or item position effects (M_3 and M_4) indicated significantly better model-data fit than the Rasch model (M_1), whereas the SEM model with form effects (M_2) was not better than the Rasch model in terms of model-data fit. This finding is not surprising given the non-significant form effects in M_2. Among the four models that included additional predictors, M_2 with form effects indicated the worst model fit, whereas M_4 with only significant item position effects indicated the best model fit based on both AIC and BIC.

Simulation study

The empirical study demonstrated that the SEM approach is capable of detecting different types of position effects in large-scale assessments. In the second part of this study, the Monte Carlo simulation capabilities of Mplus (Muthén and Muthén 1998–2015) were used to investigate the extent to which the proposed SEM approach can detect item position effects accurately. More specifically, the recovery of the SEM model in the presence of a linear position effect on item difficulty was evaluated. Hit rates in detecting items with the linear position effect and Type I error rates in flagging items with no position effects were examined via simulated data sets.

Simulation design

To model the item characteristics of a large-scale assessment, item difficulty parameters obtained from the Rasch model in the empirical study were used as the population

Table 6 Summary of the model fit information from the SEM models

Model	Number of parameters	Loglikelihood value	Scaling factor	AIC	BIC
M_1: Rasch model	45	−257,591	0.992	515,272	515,604
M_2: Form effect	48	−257,590	1.008	515,277	515,631
M_3: Passage position effect	52	−257,571**	1.007	515,246	515,629
M_4: Item position effect	90	−257,420**	0.991	515,019	515,683
M_4: Item position effect*	68	−257,437**	0.992	515,010	515,511

* This model only included the significant item position effects in M_4

** The model fits the data significantly better than M_1 at the alpha level of $\alpha = .05$

parameters when generating item responses in the simulation study. The item difficulty parameters ranged from −2.894 to 0.196 ($M = -1.22$, $SD = 0.63$). Examinees' latent traits were drawn from a standard normal distribution, $\eta \sim N(0, 1)$. Two test forms with 45 dichotomous items were created. Six items were manipulated to have linear position effects and the difficulty of the remaining items were not influenced by position changes. Table 7 shows how item difficulty parameters of the six items were manipulated as a result of their positions across the two test forms. There was no correlation between item positions and the latent trait because item positions were randomly determined independent of the latent trait.

The simulation study consisted of three factors: (a) the magnitude of the linear position effect on item difficulty (.01 and .02 per one position change); (b) sample size (1000, 5000, and 10,000 examinees); and (c) the size of position change for the six manipulated items across two forms (+10, +25, and +40 positions). The chosen values for the linear position effect were similar to those found in the empirical study as well as the position effects reported in earlier studies (e.g., Debeer and Janssen 2013). Similarly, sample size values resemble the number of examinees from previous empirical studies on item position effects (e.g., Bulut et al. 2016; Debeer and Janssen 2013; Qian 2014; Weirich et al. 2016). For each crossed condition, 1000 data sets were generated.

Model estimation

Each simulated data set was analyzed in Mplus using the SEM model in Eq. 7. Hit rates (i.e., correctly detecting items with a linear position effect) and Type I error rates (i.e., falsely flagging items with no position effects) were evaluated. Furthermore, recovery of item difficulty parameters was studied descriptively with root mean squared error (RMSE) and mean error (ME) as follows:

$$RMSE = \frac{1}{K}\left(\sum_{k=1}^{K}\sqrt{\frac{\sum_{i=1}^{n}(b_i - \hat{b}_i)^2}{n}}\right), \tag{10}$$

$$ME = \frac{1}{K}\left(\sum_{k=1}^{K}\frac{\sum_{i=1}^{n}(b_i - \hat{b}_i)}{n}\right), \tag{11}$$

Table 7 Summary of the items with linear position effects in the simulation study

Position in form 1	Position in form 2	Change in position	Position effect per one position	Total position effect	Item difficulty in form 1	Item difficulty in form 2
1	41	40	−.01	−.4	−0.99	−0.59
2	27	25	−.01	−.25	−0.98	−0.73
3	13	10	−.01	−.1	−0.97	−0.87
4	44	40	−.02	−.8	−0.92	−0.12
5	30	25	−.02	−.5	−0.90	−0.40
6	26	10	−.02	−.2	−0.88	−0.68

where b_i is the true value of item difficulty of item i, $\hat{b}_i$ is the estimated value of item difficulty for item i, and n is the total number of items (i.e., 45), and K is the number of replications (i.e., 1000).

Results

Table 8 shows the results of the simulation study. The 95% coverage indicates the proportion of replications for which the 95% confidence interval contains the population value of the linear position effect. Hit rate is the proportion of replications for which the items with position effects were correctly flagged for exhibiting a linear position effect at the $\alpha = .05$ level. Type I error rate is the average proportion of replications for which the items with no position effects were falsely flagged for exhibiting a linear position effect at the $\alpha = .05$ level. For every crossed condition, the estimated values of the position effect were very similar to the population values of the position effect. Furthermore, the 95% coverage column in Table 8 shows that the 95% confidence intervals of the estimated position effects covered the population values of the position effects 94% or more of the time. These findings suggest that the SEM model could successfully recover the position effect parameters, even when the size of the position effect was small and sample size was not large.

Hit rates appeared to vary depending upon sample size, the magnitude of the position effect, and the size of position change between two forms. When the magnitude of the position effect was larger ($\beta = -.02$ per one position change), hit rates were very high, except for the condition where sample size was 1000 and the size of position change between two forms was 10. Hit rates improved as sample size, the magnitude

Table 8 Hit rates and Type I error rates in the simulation study

Simulation factors				Estimated position effect				Estimated item difficulty	
Sample size	Position effect	Change in position	Total position effect (β)	$\hat{\beta}$	95% coverage	Hit rate	Type I error rate	RMSE	ME
1000	−.01	40	−.4	−.405	.958	.723	.051	.0058	.0033
	−.01	25	−.25	−.248	.951	.351			
	−.01	10	−.1	−.097	.951	.086			
	−.02	40	−.8	−.805	.947	.999			
	−.02	25	−.5	−.504	.957	.892			
	−.02	10	−.2	−.204	.948	.252			
5000	−.01	40	−.4	−.402	.951	1	.049	.0017	.0007
	−.01	25	−.25	−.249	.953	.945			
	−.01	10	−.1	−.099	.946	.279			
	−.02	40	−.8	−.8	.94	1			
	−.02	25	−.5	−.499	.96	1			
	−.02	10	−.2	−.203	.96	.821			
10,000	−.01	40	−.4	−.402	.947	1	.049	.0009	<.0001
	−.01	25	−.25	−.248	.95	.999			
	−.01	10	−.1	−.099	.951	.48			
	−.02	40	−.8	−.8	.942	1			
	−.02	25	−.5	−.498	.961	1			
	−.02	10	−.2	−.202	.954	.976			

of the magnitude of the position effect, and the size of position change between two forms increased. Hit rates for the condition in which the magnitude of the position effect was −.01 per one position change and the size of position change was 10 remained low, despite increasing sample size from 1000 to 10,000. The average Type I error rates for the items with no position effects were near the nominal rate ($\alpha = .05$), which indicates that the SEM model did not falsely flag items for exhibiting position effects. The RMSE and ME values for the estimates of item difficulty were quite small, suggesting that the recovery of the item difficulty parameters was good. The size of the RMSE and ME values decreased even further as sample size increased.

Discussion and conclusions

Item position effect, which often is viewed as a context effect in assessments (Brennan 1992; Weirich et al. 2016), occurs when the difficulty or discrimination level of a test item varies depending on the location of the item on the test form. For example, the difficulty of an item can increase in later positions due to a fatigue effect or decreasing test-taking effort (Hohensinn et al. 2011; Weirich et al. 2016). To investigate item position effects, researchers have proposed different approaches using logistic regression (e.g., Davey and Lee 2011; Pomplun and Ritchie 2004), multilevel IRT models based on the GLMM framework (e.g., Albano 2013; Li et al. 2012; Weirich et al. 2014), and test equating (e.g., Pommerich and Harris 2003; Meyers et al. 2009; Store 2013). The purpose of the current study was to introduce a factor analytic approach for modeling item position effects using the SEM framework. In the first part of the study, the methodological capabilities of the proposed SEM approach were illustrated in an empirical study using data from an operational testing program in reading. Test form, passage position, and item position effects were investigated. In the second part of the study, a Monte Carlo simulation study was conducted to evaluate the accuracy of the SEM approach in detecting item position effects. The simulation study showed that the SEM approach is quite accurate in detecting linear item position effects, except for the conditions in which both the number of examinees and the magnitude of the item position effect are small.

The proposed SEM approach contributes to the literature of item position effects in large-scale assessments in three ways. First, the SEM approach allows researchers and practitioners to examine both linear position effects and interaction effects in the same model. It is typically assumed that item difficulty linearly increases or decreases as the items are administered in later positions. However, changing item positions can also result in changes in item difficulty as a result of the interaction of item positions and the latent trait (e.g., Debeer and Janssen 2013; Weirich et al. 2014). Hence, it is important to evaluate both linear position effects and interaction effects when designing a large-scale assessment containing multiple forms with different item orders or with randomized item ordering. Second, the SEM approach presented in this study is a flexible method for studying position effects with various IRT models for dichotomously and polytomously scored items—such as the two-parameter model, Partial Credit Model, and Graded Response Model. For example, the position analyses in the empirical part of this study could be easily extended to the two-parameter IRT model by freely estimating factor loadings of the items (see Fig. 2). Third, the SEM approach is applicable to large-scale assessments with more complex designs, such as multiple test forms (or booklets) consisting of the same set of items in

different positions, test forms with completely randomized item ordering for each examinee, and multiple matrix booklet designs (Gonzalez and Rutkowski 2010).

Significance and future research

The current study has important implications in terms of educational testing practices. First, this study evaluated position effects in a large-scale assessment. The proposed SEM approach can help practitioners identify problematic test items with significant position effects and thereby leading to large-scale assessments with improved test fairness. Second, this study presents a straightforward and efficient approach to investigate different types of position effects (e.g., item position effect, passage position effect, and form effect). Hence, the proposed approach can be easily applied to assessments with a large number of items and examinees. Third, the results of this study can provide guidance for further research on position effects in computer-based and computerized adaptive tests. For example, future research can focus identifying which types of items are more likely to exhibit position effects in computer-based assessments and computerized adaptive tests. This can help practitioners select the most appropriate items when designing computer-based assessments and computerized adaptive tests.

This study introduced the SEM model that incorporates interaction effects, but did not investigate the statistical properties of the proposed model. Thus, further research is needed to evaluate the adequacy and accuracy of the proposed SEM model in detecting interaction effects. Given the increasing popularity of multidimensional IRT models, it would also be worthwhile to evaluate position effects in large-scale assessments that measure multiple latent traits. Finally, as Debeer and Janssen (2013) pointed out, there is a lack of research on the underlying reasons of item position effects in large-scale assessments. Future research with the SEM approach can include item-related predictors (e.g., cognitive demand, linguistic complexity) and examinee-related predictors (e.g., gender, test motivation, anxiety) to explain why item position effects occur in large-scale assessments.

Authors' contributions

OB and MJG developed the theoretical framework for the analysis of position effects using structural equation modeling. Furthermore, OB and QG carried out simulation and real data analyses for the study. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Appendix 1: Mplus code for testing passage position effects

Appendix A

Mplus Code for Testing Passage Position Effects

```
TITLE: Model to examine passage position effects
DATA: FILE IS position.csv;
!i1-i45 are items; pp1 to pp7 are positions for seven passages
VARIABLE: NAMES ARE i1-i45 pp1-pp7;
CATEGORICAL ARE i1-i45;
ANALYSIS: ESTIMATOR = MLR;
MODEL: f BY i1-i45@1;
f@1;

!Each passage is linked to a group of items;
i31 i32 i33 i34 i35 i36 i37       ON      pp1 (a);
i9 i10 i11 i12 i13 14             ON      pp2 (b);
i26 i27 i28 i29 i30 i41 i42 i43   ON      pp3 (c);
i3 i4 i5 i6 i7                    ON      pp4 (d);
i1 i2 i8 i38 i39 i45              ON      pp5 (e);
i21 i22 i23 i24 i25 i40           ON      pp6 (f);
i15 i16 i17 i18 i19 i20 i44       ON      pp7 (g);
```

Appendix 2: Mplus code for testing item position effects

Appendix B

Mplus Code for Testing Item Position Effects

```
TITLE: Model to examine item position effects
DATA: FILE IS position.csv;
!i1-i45 are items; p1 to p45 are positions for 45 items
VARIABLE: NAMES ARE i1-i45 p1-p45;
CATEGORICAL ARE i1-i45;
ANALYSIS: ESTIMATOR = MLR;
MODEL: f BY i1-i45@1;
f@1;
i1     ON     p1;
i2     ON     p2;
i3     ON     p3;
i4     ON     p4;
i5     ON     p5;

 ⋮     ⋮      ⋮
 ⋮     ⋮      ⋮
 ⋮     ⋮      ⋮

i40    ON     p40;
i41    ON     p41;
i42    ON     p42;
i43    ON     p43;
i44    ON     p44;
i45    ON     p45;
```

Received: 27 July 2016 Accepted: 5 February 2017
Published online: 16 February 2017

References

Abedi, J., Leon, S., & Kao, J. C. (2007). *Examining differential item functioning in reading assessments for students with disabilities*. Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment. Retrieved from https://nceo.umn.edu/docs/OnlinePubs/PARA/examiningdif/examiningDIFreport.pdf

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Albano, A. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement, 50*(4), 408–426.

Alexandrowicz, R., & Matschinger, H. (2008). Estimation of item location effects by means of the generalized logistic regression model: A simulation study and an application. *Psychology Science Quarterly, 50,* 64–74.

Arbuckle, J. L. (2011). *IBM SPSS Amos 20 user's guide*. Armonk, NY: IBM Corporation.

Asparouhov, T., & Muthén, B. (2006). *Robust Chi square difference testing with mean and variance adjusted test statistics* (Mplus Web Notes No. 10). Retrieved from https://www.statmodel.com/download/webnotes/webnote10.pdf

Asparouhov, T., & Muthén, B. (2016). *IRT in Mplus* (Technical Report). Retrieved from https://www.statmodel.com/download/MplusIRT.pdf

Bentler, P. M., & Wu, E. R. J. C. (2002). *EQS 6 for Windows user's guide*. Temple City, CA: Multivariate Software Inc.

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education, 5,* 225–264.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Bulut, O. (2015). An empirical analysis of gender-based DIF due to test booklet effect. *European Journal of Research on Education, 3*(1), 7–16.

Bulut, O., Lei, M., & Guo, Q. (2016). Item and testlet position effects in computer-based alternate assessments for students with disabilities. *International Journal of Research & Method in Education*. doi:10.1080/1743727X.2016.1262341.

Core Team, R. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Davey, T., & Lee, Y. H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised general test (Research Report 11–26)*. Princeton, NJ: Educational Testing Service.

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

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164–185.

Ferrando, P. J., Anguiano-Carrasco, C., & Demestre, J. (2013). Combining IRT and SEM: An hybrid model for fitting responses and response certainties. *Structural Equation Modeling, 20,* 208–225.

Ferrando, P. J., & Lorenzo-Seva, U. (2005). IRT-related factor analytic procedures for testing the equivalence of paper-and-pencil and internet-administered questionnaires. *Psychological Methods, 10*(2), 193–205.

Ferrando, P. J., & Lorenzo-Seva, U. (2013). *Unrestricted item factor analysis and some relations with item response theory* (Technical Report). Department of Psychology, Universitat Rovira i Virgili, Tarragona. Retrieved from http://psico.fcep.urv.es/utilitats/factor/documentation/technicalreport.pdf

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29,* 278–295.

Fox, J., Nie, Z., & Byrnes, J. (2016). *sem: Structural equation models*. [Computer software]. http://CRAN.R-project.org/package=sem

Glöckner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling, 10,* 544–565.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute Monograph, 3,* 125–156.
Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly, 50*(3), 379–390.
Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling, 54,* 418–431.
Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the linear logistic test model. *Psychology Science Quarterly, 50,* 391–402.
Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation, 17,* 497–509.
Jöreskog, K., & Sörbom, D. (2015). *LISREL 9.2 for Windows [Computer software]*. Skokie, IL: Scientific Software International, Inc.
Jöreskog, K., & Yang, F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 57–89). Mahwah, NJ: Lawrence Erlbaum.
Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15,* 136–153.
Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8,* 147–154.
Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65,* 457–474.
Klein, A., & Muthén, B. (2007). Quasi-maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research, 42,* 647–673.
Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices (3rd Edition)*. New York, NY: Springer.
Kolen, M. J., & Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement, 27,* 27–39.
Lee, S., Bulut, O., & Suh, Y. (2016). Multidimensional extension of multiple indicators multiple causes models to detect DIF. *Educational and Psychological Measurement*. doi:10.1177/0013164416651116.
Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936–949.
Li, F., Cohen, A., & Shen, L. (2012). Investigating the effect of item position in computer-based tests. *Journal of Educational Measurement, 49*(4), 362–379.
Lu, I. R. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling, 12*(2), 263–277.
MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement, 27*(5), 372–379.
Marsh, H. W., Wen, Z., & Hau, K. T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9,* 275–300.
Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international science report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: Boston College.
McDonald, R. P. (1999). *Test theory: Unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.
Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*(1), 38–60.
Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39,* 479–515.
Moosbrugger, H., Schermelleh-Engel, K., Kelava, A., & Klein, A. G. (2009). Testing multiple nonlinear effects in structural equation modelling: A comparison of alternative estimation approaches. In T. Teo & M. S. Khine (Eds.), *Structural equation modeling in educational research: Concepts and applications* (pp. 103–136). Rotterdam: Sense.
Moses, I., Yang, W., & Wilson, C. (2007). Using kernel equating to assess item order effects on test scores. *Journal of Educational Measurement, 44,* 157–178.
Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus* (Mplus Web Note No. 4). Retrieved from https://www.statmodel.com/download/webnotes/CatMGLong.pdf
Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus User's Guide Seventh Edition*. Los Angeles, CA: Muthén & Muthén.
Pommerich, M., & Harris, D. J. (2003). *Context effects in pretesting: Impact on item statistics and examinee scores*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
Pomplun, M., & Ritchie, T. (2004). An investigation of context effects for item randomization within testlets. *Journal of Educational Computing Research, 30*(3), 243–254.
Pritikin, J. N., Hunter, M. D., & Boker, S. M. (2015). Modular open-source software for item factor analysis. *Educational and Psychological Measurement, 75*(3), 458–474.
Qian, J. (2014). An investigation of position effects in large-scale writing assessments. *Applied Psychological Measurement, 38*(7), 518–534. doi:10.1177/0146621614534312
Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Chicago, IL: The University of Chicago Press.
Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests [Methodical challenges in the calibration of achievement tests]. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss, & G. Walther (Eds.), *Bildungsstandards Deutsch und Mathematik* (pp. 42–106). Weinheim: Beltz.
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Satorra, A., & Bentler, P. M. (2001). A scaled difference Chi square test statistic for moments structure analysis. *Psychometrika, 66,* 507–514.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.
Schweizer, K. (2012). The position effect in reasoning items considered from the CFA perspective. *International Journal of Educational and Psychological Assessment, 11,* 44–58.
Schweizer, K., Schreiner, M., & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: A two dimensional model of APM. *Psychology Science Quarterly, 51,* 47–64.
Store, D. (2013). *Item parameter changes and equating: An examination of the effects of lack of item parameter invariance on equating and score accuracy for different proficiency levels (Unpublished doctoral dissertation)*. Greensboro, NC: The University of North Carolina.
Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.
Umbach, N., Naumann, K., Hoppe, D., Brandt, H., Kelava, A., & Schmitz, B. (2017*). nlsem: Fitting structural equation mixture models.* [Computer software]. http://CRAN.R-project.org/package=nlsem
van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement, 44,* 117–130.
Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38*(7), 535–548.
Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2016). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*. doi:10.1177/0146621616676791.
Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183.
Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35,* 339–361.