Psychological Test and Assessment Modeling, Volume 60, 2018 (3), 369-385

The impact of ignoring the partially compensatory relation between ability dimensions on norm-referenced test scores

Janine Buchholz[1] *& Johannes Hartig*[2]

Abstract

The IRT models most commonly employed to estimate within-item multidimensionality are compensatory and suggest that some dimensions (e.g., traits or abilities) can make up for a lack in others. However, many assessment frameworks in educational large-scale assessments suggest partially compensatory relations among dimensions. In two Monte-Carlo simulation studies we varied the loading pattern, the latent correlation between dimensions and the ability distribution to evaluate the impact on test scores when a compensatory model is incorrectly applied onto partially compensatory data. Findings imply only negligible effects when true abilities are bivariate normal. Assuming a uniform distribution, however, analyses of differences in test scores demonstrated systematic effects for specific patterns of true ability: High abilities are largely underestimated when the other ability required to solve some of the items was low. These findings highlight the necessity of applying the partially compensatory model under data conditions likely to occur in educational large-scale assessments.

Keywords: educational testing, test interpretation, testing programs, Monte Carlo methods, validity

[1] *Correspondence concerning this article should be addressed to:* Janine Buchholz, German Institute for International Educational Research (DIPF), Frankfurt am Main, Schloßstraße 29, 60486 Frankfurt, Germany, email: buchholz@dipf.de

[2] German Institute for International Educational Research (DIPF)

Educational large-scale assessments (LSAs) such as the *Programme for International Student Assessment* (PISA), *Trends in International Mathematics and Science Study* (TIMSS), and *Progress in International Reading Literacy Study* (PIRLS) play an important role in the national and international research and policy landscape (Rutkowski, Rutkowski & von Davier, 2014), and their number is expected to be yet increasing (Kamens & McNeely, 2010). For analyzing data from such assessments, item response theory (IRT) has become the primary tool. The individual assessment programs differ in the choice of the particular IRT model depending on the number of observed response categories and item parameters to be estimated (Berezner & Adams, 2017). Another criterion for model choice is the dimensionality of the latent construct. For example, in PISA 2009 in which reading represented the major domain of the assessment, items were developed to measure subscales representing three aspects of reading competence ('access and retrieve', 'integrate and interpret', and 'reflect and evaluate'; OECD, 2010). In addition to such multidimensionality on the test level (or '*between-item* multidimensionality': Adams, Wilson, & Wang, 1997), multidimensionality may also occur on the item level ('*within-item* multidimensionality'). In fact, most items in ability and achievement tests can be considered to be multidimensional in nature (Embretson & Yang, 2006). For example, the items of the German Educational Standards in Mathematics (KMK, 2004) are each indicative of both a content-related (e.g., 'numbers', 'space and shape') and a process-related view (e.g., 'modeling', 'communicating') of mathematical competence (Mikolajetz, 2017). For such multidimensional items, the question arises on how to integrate the multiple dimensions in order to predict the probability of success on such an item. Two respective classes of multidimensional IRT (MIRT) models can be distinguished: linear compensatory (eq. 1, in the following 'M2PL') and multiplicative partially compensatory models (eq. 2, in the following 'PC2PL'; Reckase, 2009; Sympson, 1978; Whitely, 1980).

$$P(X_{ij}=1)=\frac{e^{(a_{i1}\theta_{j1}+a_{i2}\theta_{j2}\,-\,b_i)}}{1+e^{(a_{i1}\theta_{j1}+a_{i2}\theta_{j2}\,-\,b_i)}} \tag{1}$$

$$P(X_{ij}=1)=\frac{e^{a_{i1}(\theta_{j1}-b_{i1})}}{1+e^{a_{i1}(\theta_{j1}-b_{i1})}}\times\frac{e^{a_{i2}(\theta_{j2}-b_{i2})}}{1+e^{a_{i2}(\theta_{j2}-b_{i2})}}. \tag{2}$$

In both equations, $P(X_{ij}=1)$ represents the probability of success for person j on item i in the two-dimensional case, i.e., for an item that measures two dimensions simultaneously. In equation 1, every combination of θ_{j1} and θ_{j2} that yields the same sum will lead to an identical prediction. For example, both the combinations of $\theta_{j1}=0, \theta_{j2}=0$ and $\theta_{j1}=+3, \theta_{j2}=-3$ lead to equal sums when discriminations are held constant. In the latter example, the high value of θ_{j1} makes up for the low value of θ_{j2}, thus illustrating the M2PL's compensatory nature. In the PC2PL (eq. 2), in contrast, the item is decomposed into components and a unidimensional model is applied to each of them. The overall probability of a correct response cannot exceed the maximum probability of success on one of its components. In other words, as soon as success on one part (e.g. a cognitive task in a test item) is unlikely, success on the item as a whole is modeled to be unlikely. The two models are illustrated by the surface plots in Figure 1, with the surface representing the predicted probability of success for every combination of ability levels, θ (theta). The models' predictions differ most in cases with opposing theta levels (e.g. $\theta_1 = +3$ and $\theta_2 = -3$), i.e., in the degree to which the models allow for compensation.

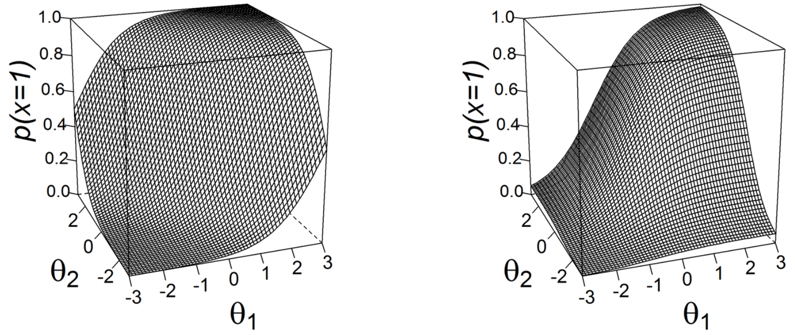


Figure 1:
Response surface plots illustrating the predicted probability of success, $P(x = 1)$, as a function of θ_1 and θ_2 for the M2PL (left; $a1 = a2 = 1.3$, $b = 0$) and the PC2PL (right; $a1 = a2 = 1.3$, $b1 = b2 = 0$).

Apart from their implications for the relation between abilities and correct responses, the models also differ in quantity and meaning of their difficulty parameter: Whereas the M2PL's parameter refers to the location in the theta space where the test item is most discriminating (Reckase, 1985), the PC2PL's parameters refer to the unidimensional difficulties for each part of the item (e.g., a cognitive task required in the solution process).

In order to yield validity evidence based on internal structure (AERA, APA, & NCME, 2014), psychometric models must reflect theoretical assumptions about the model structure. In case of a MIRT model, this requirement extends to the interplay between the multiple dimensions. As a result, a partially compensatory model should be employed as soon as multiple abilities are required *simultaneously* in order to solve an item. For example, a mathematical word problem requires the examinee to exhibit both numeric ability for solving the mathematical part and verbal ability for decoding the written text. The probability of a correct response declines as soon as one of the abilities (e.g. reading) is low, regardless of the level on the other dimension (e.g., mathematical competence). Although there are a number of assessment frameworks in educational testing in which the measured dimensions only allow for little or no compensation (e.g. KMK, 2004; NCTM, 2000; OECD, 2010; Grønmo, Lindquist, Arora, & Mullis, 2013; Jones, Wheeler, & Centurino, 2013), the compensatory M2PL is most commonly employed (Babcock, 2011). One reason might be the PC2PL's computational burden as estimation procedures have been rarely implemented in readily-available software (Babcock, 2011; Bolt & Lall, 2003).

Research interest

The two MIRT models for within-item multidimensionality described above differ with respect to their assumptions about the interplay between the dimensions of the measured constructs. A mismatch between a researcher's theoretical assumptions and the model's implications is therefore likely to affect the person parameter estimates resulting from the selected model and consequently the test scores derived from those estimates. This may lead to invalid test score interpretations and inferences. However, it appears that these effects have not been investigated yet, and we therefore aim to evaluate the magnitude of differences in persons' test scores when an existing partially compensatory relation among dimensions is ignored and a compensatory MIRT model is applied instead. Note that the focus of this study is not to investigate the recovery of true ability parameters, but to examine the *differences* of the relative position within a norm-referenced distribution of test scores derived from MIRT scaling. Such norm-referenced test scores, comparing an individual ability estimate to a population distribution, are common in both individual achievement testing and large-scale assessments. For example, test scores for intelligence are typically reported on a scale with a mean of 100 and a standard deviation of 15, and PISA scores are reported on a scale with a mean of 500 and a standard deviation of 100 across OECD countries (e.g. OECD, 2017).

We conducted two Monte Carlo simulation studies. Study 1 investigates differences in persons' test scores resulting from incorrectly applying the M2PL in an overall sample of normally distributed abilities. Study 2, in contrast, focuses on differences in persons' test scores for a specific subpopulation characterized by a certain combination of true abilities for which the models are expected to differ most (cf. Figure 1). Throughout both simulation studies, we will illustrate the findings in the context of an assessment of mathematical competence. Consider the framework provided by the German Educational Standards in Mathematics for Secondary Education (KMK, 2004) introduced above. According to this framework, persons can be described along a latent continuum on each of 11 dimensions, five dimensions representing the content-related view of mathematical competence and six dimensions representing the process-related view. Items were developed to simultaneously measure one dimension of the two views each, an example being items aiming to measure both 'space and shape' (content) and 'communicating' (cognitive process). These two subdimensions of mathematical competence have been found to correlate on a moderate level only ($\rho = .39$: Mikolajetz, 2017) which implies that there are examinees with one ability being high and one being low. Non-native speakers (English Language Learners, 'ELLs') with low (English) language proficiency ('ELP') for example, struggling with the test language, may have trouble communicating their mathematical findings despite their ability to solve mathematical problems (e.g., Martiniello, 2009). Yet, both abilities are required for solving such items, and it cannot be expected that these abilities can compensate for each other. A partially compensatory model (PC2PL) should therefore best reflect the theoretical assumptions about success on such items.

Study 1

In this first study, the impact on test scores resulting from incorrectly applying the M2PL is investigated under data conditions we consider to be typical for educational assessments. As such, we assumed a test measuring two correlated abilities that follow a bivariate normal distribution and we varied two factors across six conditions.

Method

Data generation. In a Monte-Carlo simulation study, dichotomous responses to a 45-item test were generated under a two-dimensional 2PL model containing both unidimensional indicator items for each dimension (eq. 3) and multidimensional partially compensatory items (eq. 2). Forty-five items may be regarded as a typical test length for major domains in educational LSAs. For example, the 13 booklets in PISA 2012 contained between 11 and 62 mathematics items.

$$P\left(X_{ij} = 1\right) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \tag{3}$$

Latent abilities were generated from a multivariate normal distribution $\theta \sim \mathcal{N}(\mathbf{0}, \mathbf{P})$ with variances (diagonal elements of **P)** of one and correlations between dimensions (off-diagonal elements of **P**) depending on the simulation condition (see below).

For unidimensional items (eq. 3), difficulty parameters (b_i) were drawn from a uniform distribution in the interval [-3,3], and discrimination parameters (a_i) were drawn from a log normal distribution with the parameters [ln(1),0.1]. For five multidimensional items (eq. 2), difficulty parameters were drawn from a normal distribution with $M = -1.0$ and $Var = 1.5$, and discrimination parameters were drawn from a log normal distribution with the parameters [ln(1.5),0.1] to represent well-constructed, partially compensatory items covering the whole range of abilities. We duplicated this set of five items one, three or five times, respectively, depending on the simulation condition (see below). Figure 2 displays these item parameters as well as the resulting response surfaces.

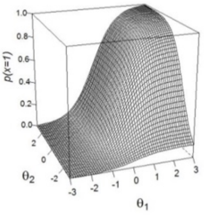
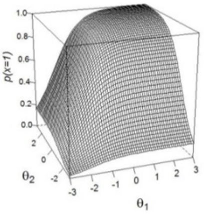
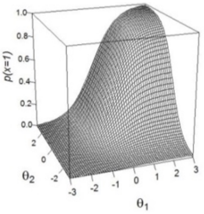
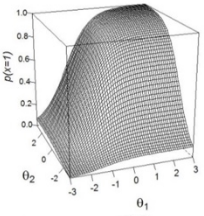
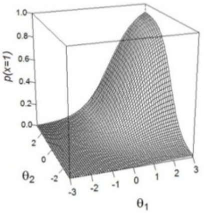
Item	Parameter				surface plot
	b_1	a_1	b_2	a_2	
1	-1.357	1.998	-1.649	0.935	
2	0.107	1.378	-2.625	1.831	
3	-2.971	1.254	-1.089	1.746	
4	-0.294	0.951	-0.083	1.100	
5	-0.485	1.919	0.446	1.888	

Figure 2.

Item difficulty (*a*) and discrimination (*b*) parameters as well as corresponding response surface plots for five multidimensional items generated under the PC2PL. This set of items was duplicated 1, 3 and 5 times for the simulation conditions with 11, 33 and 56% of items being multidimensional, respectively.

Simulation factors. We manipulated (a) the magnitude of the correlation between the two dimensions, ρ, and (b) the proportion of multidimensional items within the test, *PMI*. For (a), we expect the latent correlation to cause differences in test scores because it determines the location of cases along the two-dimensional latent space. Under high positive correlations, hardly any cases will be located in the corners of the latent space for which the two models make differential predictions (cf. Figure 1). Therefore, we varied the latent correlation to be either low ($\rho = 0.3$) or high ($\rho = 0.7$) and hypothesize that the differences in test scores between the two models are most pronounced when the latent correlation is low. More specifically, for low correlations, differences occurring under the data-generating PC2PL should be smaller than when the M2PL is estimated. When the latent correlation is high, we expect no differences between the two models. For (b), we expect that differences in test scores increase with an increase in PMI. That is, the more multidimensional items in a test, the higher the effect of the model violation should be. We manipulated PMI on the levels of 11, 33 and 56% (i.e., 5, 15 and 25 items of the 45-item test, respectively), in order to represent a wide range and allow for a thorough investigation of the pattern in differences associated with this simulation factor. The simulation design, thus, contained $2 \times 3 = 6$ conditions in total. Across conditions, sample size was held constant at $N = 2000$, and each condition was replicated 100 times.

Dependent variables. Using these simulated data, we estimated both the M2PL (eq.1) and PC2PL (eq. 2) with the Metropolis–Hastings Robbins–Monro (MH-RM) estimation algorithm (Chalmers & Flora, 2014) in the R package *mirt* (version 1.26.3; Chalmers, 2012; R Core Team, 2018). For an estimate of person ability, θ, we used EAPs with a multivariate normal prior distribution. The EAP estimates were standardized ($M = 0$, $Var = 1$) to derive norm-referenced test scores.

We prepared our findings in terms of (a) model fit and (b) differences in test scores. For (a), we calculated the percentage of replications in which the model fit indices AIC and BIC, respectively, favor the data-generating PC2PL. For (b), we used the standardized test scores and computed the squared difference between the test scores and the latent abilities used for data generation. Note that the latent abilities were also generated with $M = 0$ and $Var = 1$. Therefore, the squared differences reflect the shift of an examinee's norm-referenced test score relative to his or her true position in the sample distribution.

Results

Model estimation terminated normally across almost all conditions and replications. The proportion of replications without convergence (after 5000 iterations) varies across conditions between 0% and 3% for the M2PL, and 0% and 9% for the PC2PL, respectively. The highest proportion occurred in the condition with ρ = 0.7 and $PMI = 33$. Subsequent results are based on successful replications only. Estimation times differed by a factor of up to 6.2 in favor of the M2PL, highlighting the computational demands on estimation imposed by the PC2PL using the R package *mirt*.

Model Fit. Figure 3 illustrates the results for model fit. Across all conditions, the AIC indicated superior model fit for the data-generating PC2PL over the M2PL more often than the BIC did. More specifically, the AIC's ability to identify the correct model was best under low correlations (99 to 100%) as opposed to high correlations (60 to 73%). At the same time, its performance was rather similar across the different levels of PMI. The BIC's ability to detect the better fitting model, however, appears to be affected by both latent correlation and PMI. The BIC never identified the PC2PL over the M2PL when the latent correlation was high (0%). Under low correlations, it only detected the correct model when PMI was low (81%) or medium (25%) but never when PMI was high (0%). The weak performance of the BIC can be explained by the fact that the penalty term for the number of model parameters is larger for BIC than AIC. The more multidimensional items, the more parameters have to be estimated in the PC2PL and as a result, the better the data must be described by the model in order for the BIC to indicate superior model fit. Especially when both the latent correlation and the proportion of multidimensional items were high, the differences between the models become negligible so that the additional model parameters in the PC2PL were not necessary to better represent the data.

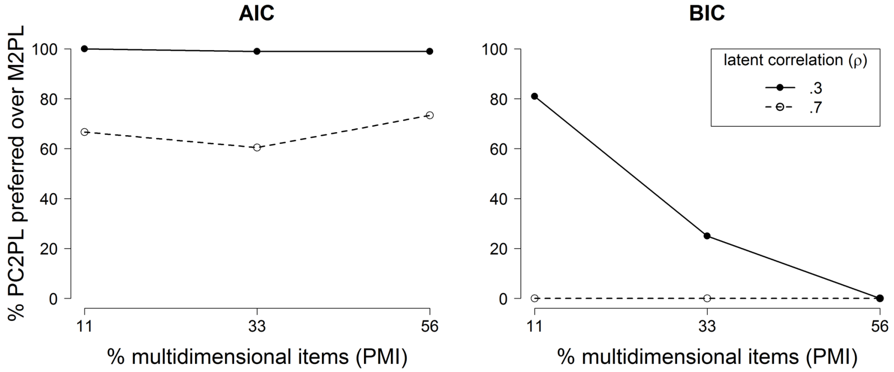


Figure 3:
Results for model fit as the percentage of replications in which AIC (left) and BIC (right), respectively, favor the PC2PL.

Differences in test scores. Table 1 shows the average squared differences between test score and true ability for both dimensions conditional on the model used and the simulation condition. Across conditions, differences resulting from the application of the data-generating PC2PL are similar or smaller than those resulting from application of the M2PL. As expected, the differences are most pronounced when either the latent correlation was low and/or the proportion of multidimensional items was high. For example, with $\rho = .3$ and $PMI = 56\%$, the mean squared difference on dimension 1 is 0.293 under the M2PL and only 0.278 under the PC2PL. When both the correlation was high ($\rho = .7$) and the proportion of multidimensional items was low ($PMI = 11\%$), the mean squared differences are about the same (0.202 on dimension 1 for both the M2PL and PC2PL).

Table 1:
Squared difference between test score $\hat{\theta}$ and true ability θ: Mean and *SE* for both models and both dimensions conditional on the simulation condition.

ρ	*PMI*	dimension 1		dimension 2	
		M2PL	PC2PL	M2PL	PC2PL
.3	11%	0.249 (0.001)	0.244 (0.001)	0.265 (0.001)	0.260 (0.001)
	33%	0.194 (0.001)	0.194 (0.001)	0.202 (0.001)	0.202 (0.001)
	56%	0.204 (0.001)	0.194 (0.001)	0.273 (0.001)	0.261 (0.001)
.7	11%	0.161 (0.001)	0.162 (0.001)	0.198 (0.001)	0.202 (0.001)
	33%	0.204 (0.001)	0.192 (0.001)	0.293 (0.001)	0.278 (0.001)
	56%	0.140 (0.000)	0.141 (0.000)	0.212 (0.001)	0.215 (0.001)

Note. ρ = latent correlation between dimensions, *PMI* = proportion of multidimensional items. Standard errors (*SE*) are reported in brackets. Both $\hat{\theta}$ *and* θ were *z*-standardized.

Taken together, only small differences with respect to model fit and test scores between the models existed when true abilities were distributed bivariate normal, and the effect becomes even negligible when dimensions correlate highly, regardless of the proportion of multidimensional items. In the example of the mathematics assessment introduced above, the dimensions ‘space and shape’ and ‘communicate’ correlated on a moderate level ($\rho = .39$, Mikolajetz, 2017). The biasing effect of model choice on scores for mathematical competence would not have been too strong overall, even when more than half of the items had been word problems. Since correlations in educational LSAs are often times found to be even higher than that, the study implies negligible impact of model violations on persons’ test scores.

Study 2

Study 1 demonstrated the overall impact of applying the simpler, yet incorrect compensatory model on test scores to be small to negligible. However, findings suggest a larger impact with an increase in the number of multidimensional items as well as with a decrease in the latent correlation. Under small correlations, some of the examinees are located in the corners of opposing theta levels (e.g. $\theta_1 = +3$ and $\theta_2 = -3$), for example ELLs that are highly able in mathematics. The objective of Study 2, therefore, is to evaluate the differences in test scores that result from incorrectly applying the M2PL *specifically* for those examinees in the corners of the latent space (cf. Figure 1). Again, we will frame our findings by using the example of ELLs with low ELP but high mathematical competence taking a mathematics assessment with items that require both of these two abilities simultaneously.

Method

In contrast to Study 1, our primary research interest in Study 2 is in the score differences for *specific subgroups*, conditional on their true ability level. We therefore refrained from manipulating data conditions but generated a larger sample instead. In particular, we had to make sure to generate a sufficient set of cases with high levels on one and low levels on the other dimension.

Procedure. The Monte-Carlo simulation study consists of two steps: We first analyzed response data based on bivariate normal abilities with both PC2PL (eq. 1) and M2PL (eq. 2) in order to obtain item parameters for each of the two models; we then used these item parameters to analyze the response data based on uniformly distributed abilities with both PC2PL and M2PL.

In the first step, dichotomous responses for $N = 100{,}000$ simulees taking a 45-item test were generated under the PC2PL. Figure 4 gives a schematic representation of the simulated loading pattern: Ten items each were unidimensional indicators for the two dimensions ('M' for items that measure mathematical competence, 'E' for items that measure ELP; eq. 3), and twenty-five of the items were within-item multidimensional and pursued a partially compensatory relationship among their dimensions (i.e., mathematical word problems that measure both dimensions simultaneously; 'ME'; eq. 2), thus corresponding to a PMI of 56%. We therefore used the item parameters of the corresponding simulation conditions in Study 1 (cf. Figure 2). Latent abilities were generated from a multivariate normal distribution $\theta \sim \mathcal{N}(\mathbf{0}, \mathbf{P})$ with variances (diagonal elements of **P)** of one and correlations between dimensions (off-diagonal elements of **P**) of 0.3. Study 2, therefore, corresponds to the simulation condition with $PMI = 56\%$ and $\rho = .3$ in Study 1, i.e., the condition for which the largest impact on test scores was found. These data were then analyzed under both the PC2PL (eq. 2) and the M2PL (eq. 1) using the MH-RM estimation algorithm, i.e., assuming an underlying normal distribution of thetas.

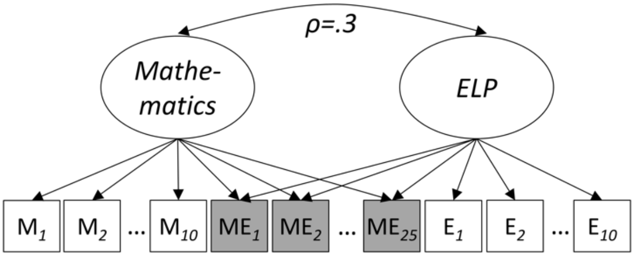


Figure 4:
Loading pattern for the data generating model (Study 2) with *M* representing items that measure mathematical competence, *E* representing items that measure ELP, and *ME* representing mathematics items that require an extensive amount of ELP.

In the second step, we simulated response data based on a bivariate uniform distribution of thetas in the interval [-3,3] and analyzed them with both PC2PL and M2PL using the item parameters obtained in the first step. By doing so, the estimation method did not rely on distributional assumptions, allowing for the estimated ability distribution to be non-normal and cover all "corners" of the latent space. Model estimation in both steps was conducted using the R package *mirt* (version 1.26.3; Chalmers, 2012; R Core Team, 2018).

Dependent variables. Data were analyzed in terms of (a) model fit and (b) differences in test scores conditional on true abilities. For (a), M2PL and PC2PL were compared with respect to AIC and BIC indices of model fit. For (b), the difference between test scores and true abilities was computed for each of the two models (M2PL, PC2PL). Just as in Study 1, these differences are based on the z-standardized EAP estimates of person ability. In order to analyze the difference in test scores conditional on the combination of the true levels of abilities, three cut points (25th, 50th, and 75th percentile rank, respectively) were applied and the results are reported separately for each of the resulting $4 \times 4 = 16$ groups of simulees. Of these, specific focus is placed on the biasing effect for simulees with high ability on one ($\theta \geq 75\%$) and low ability on the other dimension ($\theta < 25\%$), e.g., examinees with high mathematical competence and low ELP who are working on mathematical word problems. Note that these two groups ($\theta_1 < 25\%, \theta_2 \geq 75\%$ and $\theta_1 \geq 75\%, \theta_2 < 25\%$) represent examinees that rarely existed in Study 1 since the underlying true abilities were correlated. In contrast to Study 1, we calculated the simple difference between test score and true ability ($\hat{\theta} - \theta$), thus indicating whether an under- or overestimation of an examinee's relative position in the ability distribution occurred.

Results

The estimation for both steps and both models terminated successfully each. Estimation times differed by a factor of 2.3 in favor of the M2PL.

Model fit. Table 2 shows results for model fit, indicating that the PC2PL represents the data better, thus strengthening the tentative findings from Study 1. Both when the true data are bivariate normal (step 1) and when true thetas are bivariate uniform (step 2), the two indices favor the data generating PC2PL over the simpler M2PL. This is particularly encouraging since the PC2PL requires a larger number of parameters to be estimated, and the BIC penalizes models for model complexity. The result, therefore, indicates superior model fit of the PC2PL despite the model's complexity.

Differences in test scores conditional on true abilities. In general, the PC2PL led to more similar test scores over the total set of simulees, i.e., the overall span of ability levels. The squared differences between z-standardized test scores and true abilities on the first dimension are, on average, 0.141 under the M2PL and 0.094 under the PC2PL. The corresponding values on the second dimension are 0.198 and 0.142, respectively.

Table 2:
Model fit indices for compensatory (M2PL) and partially compensatory (PC2PL) model.

		M2PL	PC2PL
Step 1: bivariate normal distribution of θ	log likelihood	-2319457.093	-2317633.245
	number of parameters	116	141
	AIC	4639146.187	4635548.490
	BIC	4640249.686	4636889.812
Step 2: bivariate uniform distribution of θ using item parameters from step 1	log likelihood	-2037750.426	-2000081.347
	number of parameters	5	5
	AIC	4075510.852	4000172.695
	BIC	4075558.417	4000220.259

Note. Since the models are non-nested within each other, chi-square significance testing of the log likelihood is not possible. Since item parameters were held constant, the number of estimated parameters under the two models is equal (2 means, 2 variances and 1 covariance each).

Figures 5 and 6 give a graphic representation of differences in test scores on dimensions 1 (Figure 5) and 2 (Figure 6), respectively, that resulted from estimating either of the two models. The respective surfaces indicate the mean difference between estimated test scores and true abilities for each combination of true abilities. The PC2PL's surfaces (right) appear to be rather flat, indicating a higher degree of similarity over the whole span of ability levels whereas the surfaces resulting from the M2PL (left) are systematically skewed towards one corner, depending on the ability under scrutiny: differences in test scores on dimension 1 are largest for high levels on dimension 1 and low levels on dimension 2, and differences in test scores on dimension 2 are largest for low levels on dimension 1 and high levels on dimension 2.

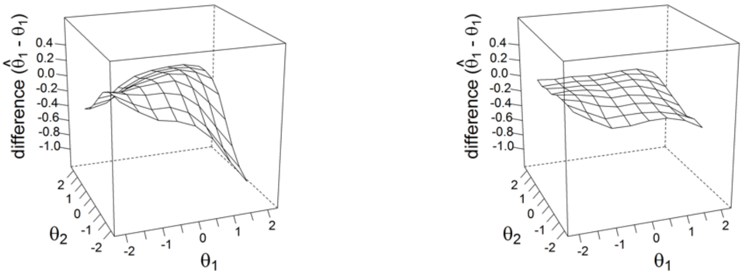


Figure 5:
Difference between test score $\hat{\theta}_1$ and true ability θ_1 under M2PL (left) and PC2PL (right) conditional on true abilities θ_1 and θ_2.

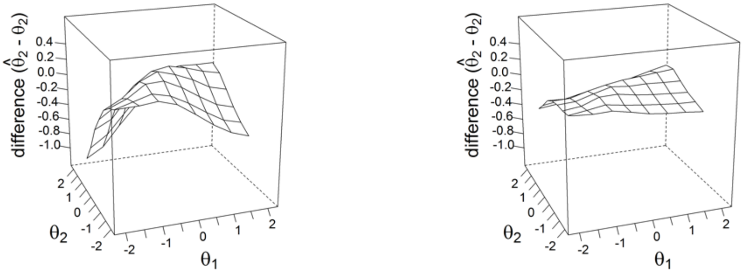


Figure 6:
Difference between test score $\hat{\theta}_2$ and true ability θ_2 under M2PL (left) and PC2PL (right) conditional on true abilities θ_1 and θ_2.

To quantify these results in greater detail, Tables 3 and 4 show the mean differences in test scores that resulted from the estimation of the M2PL (Table 3) and PC2PL (Table 4), respectively, for each of the 16 groups of combinations of true abilities. As already indicated in the surface plots, test scores obtained from the M2PL show the largest difference when the two dimensions differ most, i.e., for simulees such as ELLs highly able in mathematics. For example, with dimension 1 being high (e.g., mathematical competence, $\theta_1 \geq 75\%$ of cases) and 2 being low (e.g., ELP, $\theta_2 \leq 25\%$ of cases), the mean difference between test score and true ability $(\hat{\theta}_1 - \theta_1)$ is -0.594. Mathematical competence for ELLs, accordingly, would be underestimated by over .5 units under the standard normal distribution as a result of applying the compensatory M2PL.

Table 3:
Results for M2PL: Mean difference between test score $\hat{\theta}$ and true ability θ on dimensions 1 (top) and 2 (bottom), conditional on quartiles of true ability θ.

		Dimension 1 (e.g., mathematical competence)			
		$\theta_1 < 25$	$25 \leq \theta_1 < 50$	$50 \leq \theta_1 < 75$	$\theta_1 \geq 75$
Differences on dimension 1					
Dimension 2 (e.g.,	$\theta_2 < 25$	0.149	0.028	-0.124	-0.594
ELP)	$25 \leq \theta_2 < 50$	0.159	0.165	0.115	-0.255
	$50 \leq \theta_2 < 75$	-0.017	0.126	0.192	0.050
	$\theta_2 \geq 75$	-0.226	0.012	0.115	0.104
Differences on dimension 2					
Dimension 2 (e.g.,	$\theta_2 < 25$	0.140	0.187	-0.008	-0.226
ELP)	$25 \leq \theta_2 < 50$	0.053	0.278	0.184	0.000
	$50 \leq \theta_2 < 75$	-0.221	0.134	0.248	0.201
	$\theta_2 \geq 75$	-0.697	-0.290	-0.044	0.062

Note. Positive values indicate an overestimation, negative values an underestimation of the test score with respect to the true ability.

Table 4 shows the respective findings for differences in test scores resulting from application of the data-generating PC2PL. The mathematics test scores for simulees such as ELLs highly able in mathematics are underestimated by on average -0.145 units under the standard normal distribution. This difference is smaller compared to the one occurring when the M2PL was estimated.

Table 4:

Results for the PC2PL: Mean difference between test score $\hat{\theta}$ and true ability θ on dimensions 1 (top) and 2 (bottom), conditional on quartiles of true ability θ.

		Dimension 1 (e.g., mathematical competence)			
		$\theta_1 < 25$	$25 \leq \theta_1 < 50$	$50 \leq \theta_1 < 75$	$\theta_1 \geq 75$
Differences on dimension 1					
	$\theta_2 < 25$	0.050	-0.071	-0.055	-0.145
Dimension 2	$25 \leq \theta_2 < 50$	0.103	0.010	0.000	-0.117
(e.g., ELP)	$50 \leq \theta_2 < 75$	0.062	0.038	0.045	-0.035
	$\theta_2 \geq 75$	0.029	0.029	0.040	0.019
Differences on dimension 2					
	$\theta_2 < 25$	0.048	0.071	0.014	-0.015
Dimension 2	$25 \leq \theta_2 < 50$	0.008	0.110	0.097	0.086
(e.g., ELP)	$50 \leq \theta_2 < 75$	-0.052	0.029	0.093	0.167
	$\theta_2 \geq 75$	-0.236	-0.211	-0.154	-0.057

Note. Positive values indicate an overestimation, negative values an underestimation of the test score with respect to the true ability.

These findings generalize to the opposite situation as well: With θ_1 being low and θ_2 being high, the ability on the second dimension is underestimated as a result of applying the M2PL: The respective mean difference is -0.697 under the M2PL but only -0.236 under the PC2PL. With other words, a high level on the second dimension is underestimated as a result of the first dimension being low when the M2PL is applied.

Taken together, Study 2 showed that the relative position of examinees being high on one but low on another dimension is systematically underestimated with respect to the high ability as a result of incorrectly applying the compensatory M2PL.

Discussion

Many assessment frameworks in educational assessments suggest partially compensatory relations among the measured constructs, yet a compensatory model is most commonly employed. We therefore evaluated the impact of such model violations on estimates of person ability since misclassifications threaten the validity of test score interpretations. Although Study 1 only showed small effects overall, Study 2 demonstrated systematic effects for subgroups characterized by a specific combination of true abilities. These findings highlight the necessity of applying the partially compensatory model when there is a suspicion that one group of examinees scores low on one dimension, and the respective ability is required for some items that measure the other dimension. Otherwise, when examinees fail to solve such multidimensional items, they are systematically underestimated with respect to their (truly) high ability. Since the impact of large-scale assessments on educational policy is evident (e.g., Heynemann & Lee, 2014; Wagemaker, 2014), an inappropriate model choice may lead to invalid score interpretations about specific subgroups such as ELLs.

In the example of assessing mathematical competence in ELLs, being able to solve word problems might be considered *construct-relevant*. In this case, the researcher needs to make this explicit. In all other cases, multidimensional items such as word problems induce *construct-irrelevant* variance to the measurement of the target ability (e.g. mathematical competence). Two ways to deal with such a situation are possible: these multidimensional items are omitted from an assessment of (pure) mathematical ability, or they are modeled adequately in order to disentangle the two abilities (mathematical competence and language proficiency). The latter has been the scope of this article. Although we used the example of non-native speakers here, we want to highlight that the issue of adequately modeling multidimensional items generalizes to other examples as well.

Limitations. The two studies made some implicit and some explicit assumptions that need to be discussed. First, the PC2PL necessarily requires indicator items for each of the modeled dimensions for identification purposes. In the example of ELLs taking a mathematics assessment containing mathematics items and word problems, there also need to be items that measure language proficiency alone. Otherwise the multidimensional items cannot be modeled appropriately to account for the partially compensatory relation between the measured dimensions. In this simulation study the multidimensional items measured only two dimensions. However, the problem grows with any additional ability measured by the multidimensional items since more sets of unidimensional indicator items need to be assessed for such higher-dimensional items. This requirement might present a major limitation to some operational LSAs. Second, we generated the item discrimination, the item difficulties and the number of unidimensional indicator items to be the same for both dimensions. We thereby implicitly assumed that both dimensions are measured equally well. In operational practice, however, this might not always be the case, especially when items were constructed to be unidimensional indicators and the additional multidimensional component is more a side effect than the initial goal of the item construction process. Regarding the identical number of indicator items, only a short test of language proficiency, for example, might be assessed if it is not the actual scope of the test but intended to act as a controlling variable. As a result, language proficiency might not be measured well enough. The impact of an uneven number of unidimensional items on modeling multidimensional items, therefore, needs further inspection. Third, we generated very well discriminating items which are harder to construct and, thus, may not be as common in operational testing. However, we chose to do so in order to better investigate the pattern of test score differences. Fourth, sample size was held constant to keep the number of simulation conditions small. It can be expected that with smaller sample sizes, the pattern we found might not be as visible since even less cases would be located where the two models differ most with respect to their predicted probability of success on an item.

To date, in practically all current LSAs items are developed to be unidimensional, and multidimensional scaling models only contain between-item dimensionality. However, given the complex multidimensionality inherent in many assessment frameworks, the use of complex multidimensional models taking into account the complexity of the items appears to be a reasonable future development. The results of our studies demonstrate the importance of carefully choosing the scaling model in those settings.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23. http://dx.doi.org/10.1177/0146621697211001

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME] (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Babcock, B. (2011). Estimating a noncompensatory IRT model using Metropolis within Gibbs sampling. *Applied Psychological Measurement, 35*, 317–329. http://dx.doi.org/10.1177/0146621610392366

Berezner, A., & Adams, R. J. (2017). Why large-scale assessments use scaling and item response theory. In P. Lietz, J. C. Cresswell, K. F. Rust & R. J. Adams (Eds.), *Implementation of Large-Scale Education Assessments* (pp. 323–356). Chichester, UK: John Wiley & Sons, Ltd. http://dx.doi.org/10.1002/9781118762462.ch13

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27*, 395–414. http://dx.doi.org/10.1177/0146621603258350

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. http://dx.doi.org/10.18637/jss.v048.i06

Chalmers, R. P., & Flora D. B. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement, 38*, 339–358. http://dx.doi.org/10.1177/0146621614520958

Embretson, S. E., & Yang, X. (2006). Multicomponent latent trait models for complex tasks. *Journal of Applied Measurement, 7,* 335–350.

Grønmo, L. S., Lindquist, M., Arora, A., & Mullis, I. V. S. (2013). TIMSS 2015 mathematics framework. In I. V. S. Mullis, & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 11–27). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Heynemann, S. P., & Lee, B. (2014). The impact of international studies of academic achievement on policy and research. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 37–74). Boca Raton: CRC Press.

Jones, L. R., Wheeler, G., & Centurino, V. A. S. (2013). TIMSS 2015 science Framework. In I. V. S. Mullis, & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 29–58). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review, 54*, 5–25. http://dx.doi.org/10.1086/648471

Kultusministerkonferenz [KMK] (2004). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss [Educational standards in mathematics for second-ary education]*. Neuwied: Luchterhand.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment, 14,* 160–179. http://dx.doi.org/10.1080/10627190903422906

Mikolajetz, A. (2017). *Messung komplexer Kompetenzkonstrukte in Large-Scale-Assessments mit Hilfe von multidimensionalem adaptivem Testen* [Measurement of complex competence constructs in large-scale assessments using multidimensional adaptive testing] (Unpublished doctoral dissertation). Friedrich-Schiller-Universität Jena.

National Council of Teachers of Mathematics [NCTM] (2000). Principles and standards for school mathematics. Reston, VA: National Council of Teachers of Mathematics.

Organisation for Economic Co-operation and Development [OECD] (2010). *PISA 2009 results: What students know and can do – Student performance in reading, mathematics and science (Volume I)*. Paris, OECD.

Organisation for Economic Co-operation and Development [OECD] (2017). *PISA 2015 Technical Report*. Paris, OECD.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, London: Springer.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401–412. http://dx.doi.org/10.1177/014662168500900409

Rutkowski, D., Rutkowski, L., & von Davier, M. (2014). A brief introduction to modern international large-scale assessment. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 3–10). Boca Raton: CRC Press.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (pp. 82–98).

Wagemaker, H. (2014). International large-scale assessments: From research to policy. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 11–36). Boca Raton: CRC Press.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479–494. http://dx.doi.org/10.1007/BF02293610