Psychological Test and Assessment Modeling, Volume 54, 2012 (1), 36-53

Robustness of multidimensional analyses against local item dependence

Steffen Brandt[1]

Abstract

The negative impact of local item dependence (LID) on analyses using item response theory (IRT) has been investigated by many authors. Hitherto though, these investigations focused on unidimensional analyses. The objective of the simulation study presented here is to investigate the impact of LID on multidimensional analyses. The chosen simulation design considers tests with LID due to item bundles and compares the results of multidimensional analyses obtained with varying item bundle effect sizes, varying correlation levels for the latent traits, and different test designs. The results indicate that in multidimensional analyses LID results in a bias for the covariance estimation and that the direction of the bias interferes with the chosen test design.

Key words: test construction, item response theory, local item dependence, multidimensionality

[1] *Correspondence concerning this article should be addressed to:* Steffen Brandt, PhD, Ebereschenweg 28, 24161 Altenholz, Germany; email: steffen.brandt@artofreduction.com

An essential assumption of item response theory is local item independence; that is, beyond the variance due to one or several latent traits, the items of a test are supposed to measure, the items show no additional common variance. The negative impact of a violation of this assumption, which is denoted as local item dependence (LID), has been reported by many authors, and it has been shown that an inappropriate assumption of local item independence results in an overestimation of test information, model fit and reliability and an underestimation of the measurement error (see, e.g., Rosenbaum, 1988; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer, Bradlow, & Wang, 2007; Wen-Chung Wang & Wilson, 2005a; Yen, 1984, 1993).

It is reasonable to assume that these effects that have been investigated based on unidimensional analyses apply to the single dimensions of a multidimensional analysis in the same way. That is, without appropriate statistical modeling of existing LID their respective test information and reliability is overestimated, and the measurement error is underestimated. However, the generalization in multidimensional analyses is not that straightforward since LID can not only occur within a dimension but also across several dimensions. If, for example, two items with the same stimulus measure different constructs in a two-dimensional analysis, these two items are expected to have a higher correlation beyond that of the respective constructs they are supposed to measure. It is therefore expected that this additional correlation has an impact on the estimated covariance for the two dimensions, or more precisely, that the covariance will be overestimated due to the local item dependence. If, on the other hand, a two-dimensional analysis includes local dependence within the dimensions, it is expected that the covariance of the two dimensions is underestimated since the reliability of the respective dimension will be overestimated and the correction for the disattenuation of the covariance due to measurement error will not be appropriate.

The investigation of the impact of the effects of LID on multidimensional analyses and in particular on the corresponding covariance matrices is important since the decision on whether a data set should be interpreted unidimensional or multidimensional relies on these results. Maul (in press), for example, reanalyzed the dimensional structure of a well known measure of emotional intelligence, and found that models without consideration of LID yield a multidimensional structure; models with consideration of LID, however, did not. Wang, Cheng and Wilson (2005) investigated the impact of LID for items across tests connected by common stimuli. After applying different administration designs and models with and without consideration of LID, they found a significant impact for tests with a "parallel" design, that is, items having a common stimulus but referring to separate psychological constructs. And without consideration of LID the tests had a correlation that was .36 higher than with consideration of LID.

Despite the possible significance, as depicted by the works above, and despite the widespread application of multidimensional analyses in large-scale assessments (e.g., Martin, Gregory, & Stemler, 2000; OECD, 2002), there has not been much emphasis to date on the investigation of the possible impact of LID on multidimensional analyses. The presented simulation study, therefore, was conducted in order to depict the possible impact of LID depending on the size of the LID and the chosen administration design for a given multidimensional construct. Furthermore, the results of the simulation study provide

insight into how the differences between the results with and without consideration of LID arise.

Preparatory considerations for the design of the simulation study

A typical source of LID are item bundles. An item bundle is a set of items (also denoted as testlet) that is linked to a common stimulus (cf. Wainer & Kiely, 1987). The common stimulus for these items usually results in local item dependencies referred to as item bundle effect. The actual impact of such LID on multidimensional analyses depends on a large variety of factors: the number of dimensions, the numbers of items per dimension, the numbers of items in each item bundle, the correlations between the dimensions, the extents of the item bundle effects, and the extents of the variances of the single dimensions. A simulation study considering just two different conditions for each of these factors will result in a total of 128 different conditions for the overall test. In order to reduce the amount of test conditions, it was therefore chosen to generate different conditions on the bases of an exemplary, given multidimensional construct with a fixed amount of items and item bundles, and only the extents of the item bundle effects, the correlations of the dimensions, and the test design characteristic (see description below) are varied.

The chosen multidimensional construct roughly follows the structure of the mathematics achievement test of the Programme for International Student Assessment (PISA) 2003 (OECD, 2005). The PISA mathematics achievement test comprises four dimensions: Quantity, Change and Relationships, Space and Shape, and Uncertainty. Each dimension is measured by 20, 22, 20, and 22 items respectively. Seventy-six of these eighty-four items are dichotomous, seven have three score categories, and one has four score categories. Forty-two of these items were administered within item bundles. In order to give an impression of the extents of the item bundle effects in the real data set, the extents of the item bundle effects were investigated by a reanalysis for the German subsample using the Rasch testlet model (Wen-Chung Wang & Wilson, 2005a; 2005b; see description below). The Rasch testlet model is a restricted hierarchical model (Holzinger & Swineford, 1937; Li, Bolt, & Fu, 2006) that bases on the testlet model by Bradlow, Wainer, and Wang (1999) and is an extension of the standard Rasch model (Rasch, 1980) by an additional parameter which describes the interaction between persons and items within an item bundle. Wang and Wilson denote this parameter $\gamma_{nd(i)}$, representing the interaction between person n (n=1,..., N, and N the number of persons) and item i within item bundle $d(i)$ (i=1,..., I, and I the number of items; $d(i)$=1,..., D, and D the number of item bundles). The model equation is

$$\log(p_{ni1}/p_{ni0}) = \theta_n - b_i + \gamma_{nd(i)}, \qquad (1)$$

where p_{ni1} and p_{ni0} are the probabilities of scoring 1 and 0 on item *i* for person *n*, respectively, θ_n is the ability of person *n*, and b_i is the difficulty of item *i*. For the identification of the model several constraints have to be applied. In order to fix the locations of the scale for the latent trait and those for the item bundle effects, the means of all dimensions are set to zero. For rotational invariance the covariances of the dimension θ_n with the dimensions for the item bundle effects are set to zero. Furthermore, the item bundle effects themselves are assumed to be independent to each other.

The results of the analysis using the Rasch testlet model are given in Table 1. They show that the effects differ strongly across item bundles and range from 0.34 to 2.94, with an average variance of 1.20 for the item bundle effects and a variance of 1.95 for the measured overall mathematics achievement.

Besides the extents of the item bundle effects, the used test design plays an important role for the impact of the local item dependencies. Following the terminology of Wen-Chun Wang et al. (2005), possible test designs for multidimensional constructs are sequential and parallel test designs. These two test designs are exemplarily depicted in

Table 1:
Calibration results for the German subsample of the mathematics achievement test of PISA 2003 using the Rasch Testlet Model

Dimension	Items	Variance
Mathematics achievement	1-84	1.95
Bundle 1	3, 4, 5, 6	1.43
Bundle 2	11, 12, 13	2.94
Bundle 3	21, 22	1.09
Bundle 4	23, 24, 25	0.34
Bundle 5	26, 27, 28, 29	0.52
Bundle 6	31, 32, 33	1.61
Bundle 7	34, 35	1.36
Bundle 8	36, 37	0.75
Bundle 9	39, 40	0.45
Bundle 10	47, 48, 49	0.54
Bundle 11	51, 52	0.68
Bundle 12	64, 65, 66	1.87
Bundle 13	70, 71	2.76
Bundle 14	73, 74, 75	0.38
Bundle 15	78, 79	0.87
Bundle 16	82, 83	1.61

Figure 1 for a three-dimensional construct comprising six item bundles with three items each. In the sequential test design on the left, each dimension comprises six items from two different item bundles. In the parallel test design on the right, each dimension comprises as well six items but from six different item bundles. That is, in the first case each item bundle measures only a single dimension, whereas in the latter each item bundle measures all three dimensions. An uncountable number of other multidimensional test designs that are mixtures of the parallel and the sequential test design are possible. However, the parallel and the sequential test designs can be considered as the extremes of these possible test designs. In order to investigate the full range of the possible impact on multidimensional analyses, it is therefore useful to consider the results of a simulation study for these extremes. Additionally, parallel and sequential test designs are common in test construction. A well known test using a parallel design, for example, is the multidimensional Self-Description Questionnaire III by Marsh and O'Neill (1984). An example for a sequential test design is given by the PISA study, in which the domains mathematics, reading, and science are measured with item bundles which are entailing items from one distinct dimension only (cf. OECD, 2005).

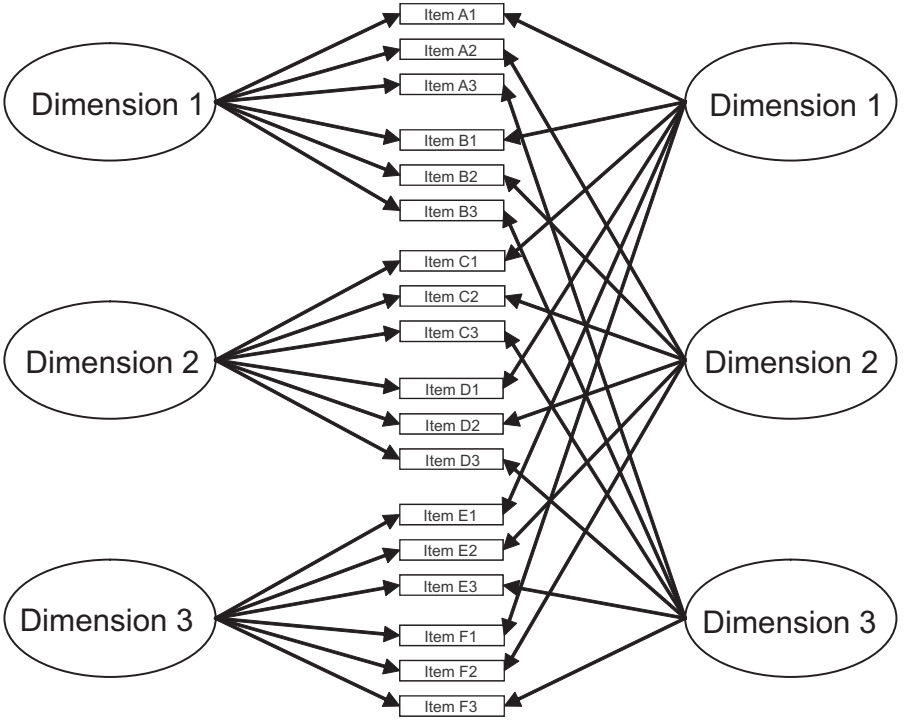

Figure 1:
Depiction of a three-dimensional sequential test design (on the left) and a three-dimensional parallel test design (on the right).

Simulation study design

The following three test characteristics are varied in the conducted simulation study: (a) the test design, (b) the extent of LID due to item bundles, and (c) the correlations between the measured dimensions.

For the above given reasons, the considered test designs are a test with item bundles in a sequential test design, a test with item bundles in a parallel test design, and additionally a reference test without item bundles, that is, without local item dependencies. Following the definition of small, medium, and large item bundle effects given by Wen-Chung Wang and Wilson (2005b), variances of 0.5, 1.0 and 2.0 for the item bundle effects are considered with variances of 2.0 for the measured latent traits. The considered levels of correlations between the latent traits are .5, .7, and .9, representing medium to high correlations and were chosen according to the extents of correlations that are typically observed in multidimensional constructs. In particular, the consideration of a correlation of .9, therefore, does not question whether such dimensions might in fact be unidimensional or not but solely corresponds to regularly reported correlations (cf. Martin et al., 2000; OECD, 2005).

As previously mentioned the used multidimensional construct follows that of the mathematics achievement test of PISA 2003. Some test characteristics are modified though, in order to allow for a more lucid presentation of the results. The number of items is adjusted to be 20 per each dimension, and each item is assigned to an item bundle of four items (according to the given test design), resulting in a total of 20 item bundles for the test. The item difficulty parameters are taken from a unidimensional, dichotomous[2] reanalysis of the German PISA 2003 mathematics achievement data and range from -2.79 to 2.56 except for one item with a difficulty of 4.07. Further, the variances for the four dimensions are set to 2.0 with a mean ability of zero. The variance of 2.0 was chosen in order to consider a scale close to that of the empirical data; it coincides with the estimated variance for the above presented example using the Rasch testlet model.

Data generation and analysis

The data generation is based on a multidimensional extension of the Rasch testlet model by Wang and Wilson (2005a; 2005b). The single steps in order to generate the simulation data according to the model are as follows (cf. Wen-Chung Wang & Wilson, 2005b):

1. Person parameters for the four-dimensional multivariate distribution are generated using ConQuest (Wu, Adams, & Wilson, 1998).
2. Normally distributed variables representing the item bundle effects are generated using SPSS for Windows.

[2] Only answers in the highest score category received a score of 1; all other answers received a score of 0.

3. The generated person parameters (θ) and random variables (γ_k), as well as the predefined item parameters (b) are used to calculate the corresponding answer probabilities using Equation 1.
4. The calculated answer probabilities are compared to a random number from the uniform [0, 1] distribution, and the simulated item response is defined as 1 if the random number is less than or equal to the associated probability, and 0 otherwise.

For each of the 21 test conditions with the characteristics given in the previous section, one hundred data sets with 1000 cases each are generated. Each data set is analyzed using the unidimensional Rasch model, the multidimensional Rasch model (Adams, Wilson, & Wang, 1997; Rost, 1996), and the Rasch subdimension model (Brandt, 2008, 2010). While the analyses using the multidimensional model show the extents of the observed bias due to the generated item bundle effects, the analyses using the subdimension model show the origin of the observed bias. The results of the unidimensional model were included as reference in order to depict possible biases on decisions on the dimensionality of data sets.

The unidimensional Rasch model coincides with the above given Rasch testlet model without the extension by the parameters γ. That is, the model equation is given by

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

For the given test data, the multidimensional model applied for the analysis can be expressed as the multicategorical, multidimensional Rasch model (Rasch, 1961; cf. Rost & Carstensen, 2002):

$$\log(p_{ni1}/p_{ni0}) = \theta_{nd} - b_{id} \,, \tag{3}$$

where θ_{nd} is the ability of person n for dimension d, and b_{id} is the difficulty of item i for dimension d.

The additionally applied subdimension model corresponds to a modified hierarchical model (Holzinger & Swineford, 1937), in which each item loads on a general factor, in the context of the subdimension model referred to as main dimension, and a specific factor, in the subdimension model referred to as subdimension. In contrast to the simple hierarchical model, however, the specific factors (subdimensions) are assumed to correlate. The definition of the subdimension model is given by

$$\log(p_{ni1}/p_{ni0}) = \theta_n + \gamma_{nd(i)} - b_i \,, \tag{4}$$

where $d(i)$ is defined as the subdimension of item i; $\gamma_{nd(i)}$ is the strength or weakness of person n in subdimension $d(i)$ relative to its ability in the main dimension; and p_{ni1}, p_{ni0}, θ_n, and b_i are defined as above (cf. Brandt, 2008). For the identification of the model, the person parameters are constrained to a mean of zero, and the covariance between the main dimension and the subdimensions is set to zero, which is common to all hierarchical models. However, in contrast to the testlet model the covariances between the subdimen-

sions are not constrained to zero but the sum of the specific abilities for each person is constrained to zero; that is, $\sum_d \gamma_{nd} = 0$ for all $n = 1, \ldots, N$. For the analyses of the simulation data, the subdimensions correspond to the dimensions of the four dimensions of the multidimensional construct, while the main dimension represents the general factor measured commonly by these four dimensions.

All considered models are special cases of the multidimensional random-coefficients multinomial logit model (MRCMLM; Adams et al., 1997) and can therefore be estimated using ConQuest (Wu et al., 1998). The unidimensional estimations were conducted using the Gauss-Hermite quadrature integration method with 100 nodes. Due to their higher complexity the multidimensional and the subdimensional estimations were estimated using the Monte Carlo integration method with 4000 nodes. The convergence criterion for all estimations was 0.01 for the change in parameters.

The estimation results of the models are compared with regard to their deviances (-2 log likelihoods)[3], their correlations, and their variances for the given four-dimensional construct depending on the extent of the generated correlation, the size of the generated item bundle effect, and the chosen test design.

Results

In order to facilitate a lucid presentation, the results of the 100 calibrated data sets per test condition as well as the dimensions' variances and correlations (which were generated to be equal) were summarized calculating their means. The results of the unidimensional model were not considered separately for the sequential and the parallel test design since the model yields equal results for both test designs.

Figure 2 depicts the changes in deviance for the unidimensional and multidimensional analyses in dependence of the size of the item bundle effects. With generated higher correlations between the four dimensions the fit of the unidimensional model is, as expected, closer to that of the multidimensional model. Furthermore, for all analyses the model fit decreases (i.e., the deviance increases) with increasing item bundle effects. However, for the multidimensional analyses the magnitude of the decrease in model fit depends on the chosen test design. In the presence of item bundle effects, the data according to a sequential test design yields a better model fit than the data according to a parallel test design. Therefore, the difference between the fit of an unidimensional model and a multidimensional model is not only affected by the correlation between the considered dimensions but as well by the size of the item bundle effects and the chosen test design. The relatively stronger decrease in model fit for the parallel test design is particularly notable for the case of large item bundle effects with generated correlations of .9, in

[3] Results for model fit indices such as Akaike's information criterion (Akaike, 1974) or the Bayesian information criterion (Schwarz, 1978) are not reported since the observed differences in the deviances are of such extent that the criteria do not provide additional information.

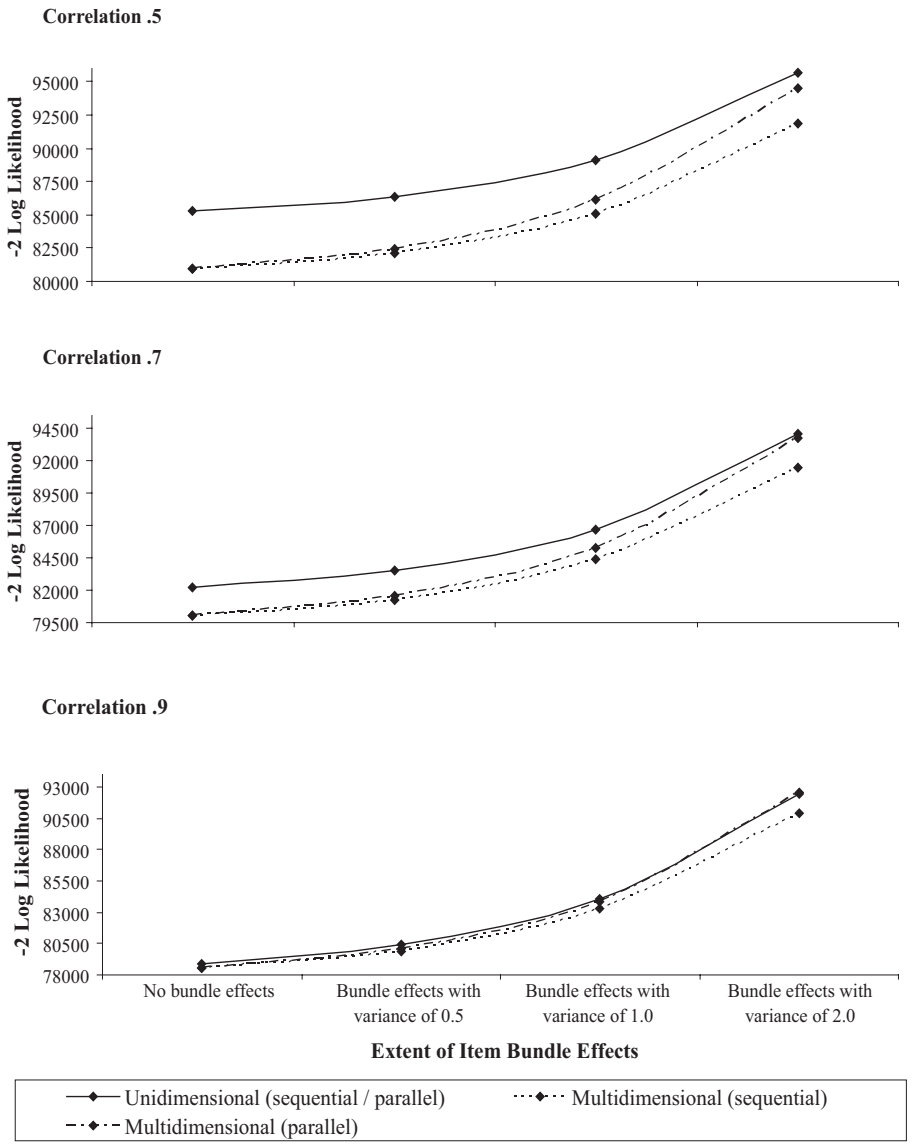

Figure 2:
Likelihoods for the unidimensional and multidimensional estimations of the four-dimensional construct with correlations of .5, .7, and .9; a sequential or a parallel test design; and varying extents of item bundle effects.

which the model fit of the unidimensional model in fact exceeds that of the multidimensional model. Here, the average deviance for the multidimensional model is 92559.5 while the deviance of the unidimensional model is 92418.8 (cf. Table A3 in the Appendix). For the sequential test design, however, the magnitude of the difference in model fit between the unidimensional and the multidimensional model seems to be comparatively independent of the size of the item bundle effects.

Figure 3 depicts the change of the estimated correlations of a multidimensional calibration depending on the generated correlations, the size of the item bundle effects, and the chosen test design. Corresponding to the model fit, the extents of the estimated correlations for the four-dimensional construct depend on the size of the item bundle effects and the chosen test design. The differences between the estimated correlations in dependence of the chosen test design are very similar for all three generated correlations. For small, medium, and large item bundle effects, the sequential test design on average results in correlations that are 0.02, 0.07, and 0.23 lower than for the parallel test design (cf. results in Appendix B). While these differences solely depend on the extent of the item bundle effect, the biases of the respective estimates in comparison to the originally generated correlations depend on the level of the generated correlation. While the bias is of equal magnitude for a generated correlation of .5, for a correlation of .9 only the parallel test design shows bias and the parallel test design seem to provide unbiased correlation estimates, independently of the size of the item bundle effects.

In order to consider the origins of the differing biases in more detail, the results of the subdimension model are considered. The model allows separating the variance components of a multidimensional construct into the common variance component (responsible for the correlation of the dimensions) and the dimension-specific variance component.

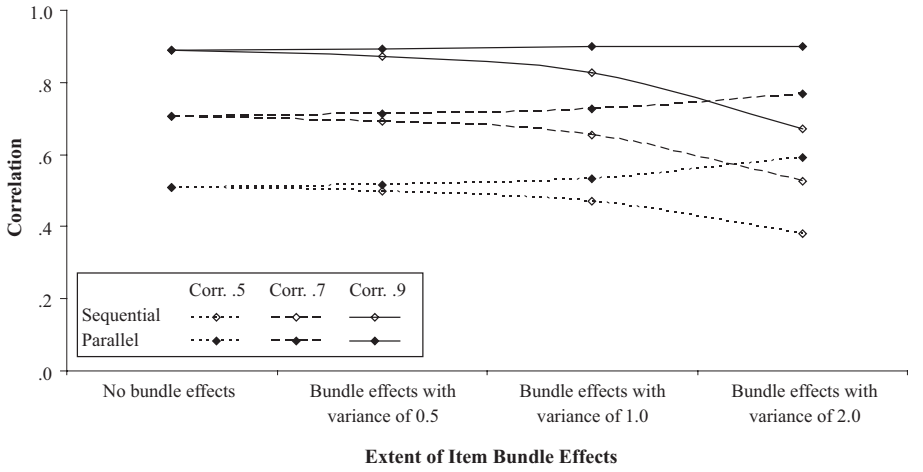

Figure 3:
Estimated correlations for the four-dimensional construct with generated correlations of .5, .7, and .9; a sequential or a parallel test design; and varying extents of item bundle effects.

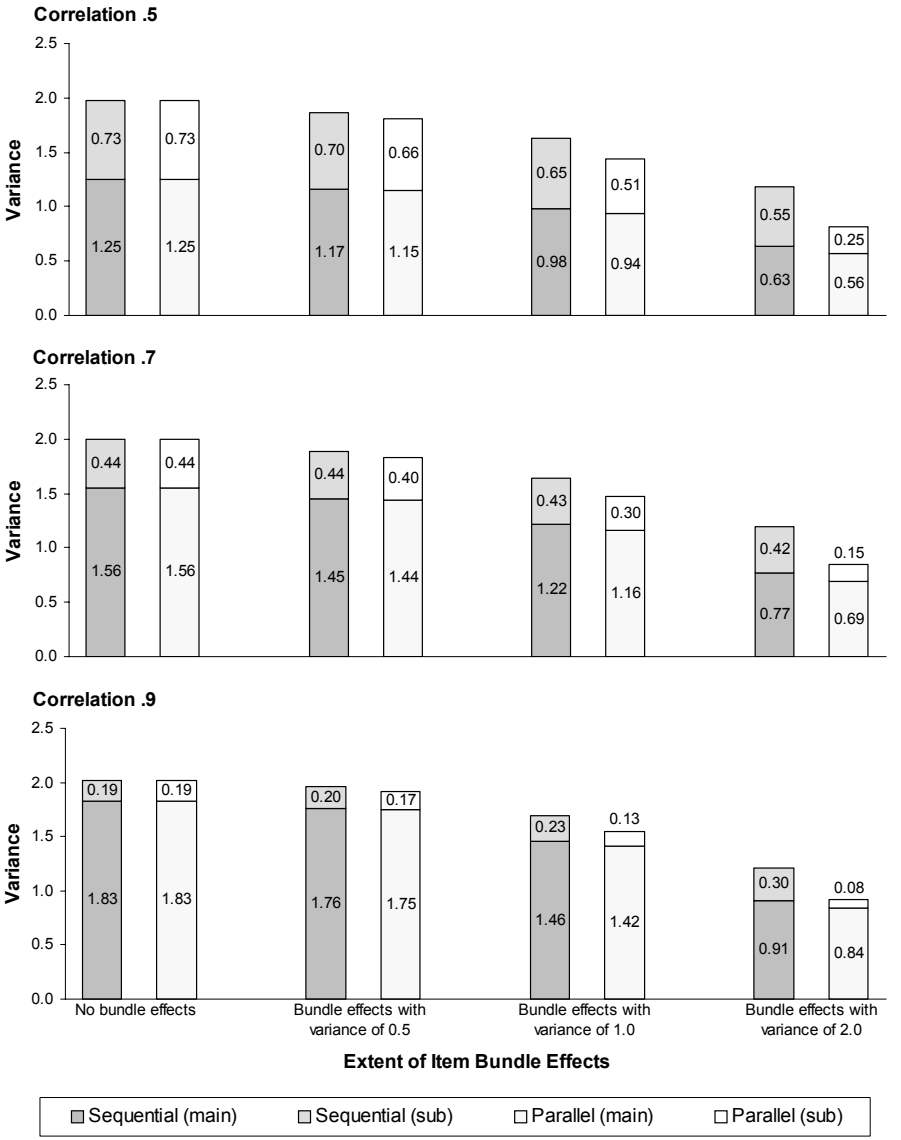

Figure 4:
Depiction of the dimensions' unidimensional (main dimension = main) and dimension specific (subdimension = sub) variance components depending on the generated correlations, a sequential or parallel test design, and the extent of item bundle effects.

The variance components corresponding to the results presented in Figure 3 are depicted in Figure 4. Here, larger proportions of unidimensional common variance within the estimated total variance represent higher correlations. Figure 4 demonstrates that for all considered test conditions the estimated total variance decreases with increasing item bundle effects. The common unidimensional variance component and the dimension-specific variance component are affected differently, though. For the sequential test design with generated correlations of .7, for example, the absolute variance of the dimensions-specific variance component stays almost unchanged (0.44 to 0.42 for no to large item bundle effects) while the common variance components decreases from 1.56 to 0.77 (again for no to large item bundle effects). Considering the differences between the sequential and the parallel test design, the item bundle effects lead to a larger decrease in total variance when using a parallel test design. The difference in the decrease of the total variance is mainly attributable to the dimension-specific variance components, however. While the unidimensional variance components differ at most at 0.08, the differences for the dimension-specific variance components extend to 0.30. The difference between the two test designs gets most visible for correlations of .9 and large item bundle effects. Here, the dimension-specific variance component is more than three times as large for the sequential test design as for the parallel test design (0.30 vs. 0.08). Furthermore, a comparison with the result of the calibration without item bundle effects shows that for correlations of .9 the introduction of item bundle effects results for the sequential test design in an increase of the dimension-specific variance beyond the dimension-specific variance that was originally generated. While the generated dimension-specific variance is 0.19, the introduction of item bundle effects results in dimension-specific variances of .20, 0.23, and 0.30 for small, medium, and large item bundle effects, respectively.

Discussion

For a better understanding of the results, it is necessary to recall the origin of the two test designs. By assigning each item of an item bundle to the same dimension, each item bundle in a sequential test affects a particular single dimension. In the parallel test design, on the other hand, each item of an item bundle loads on a different dimension. From a single dimension's perspective, therefore, not an item bundle is added to the dimension but just a single item. That is, here, the dimensions do not include item bundles in its actual sense. And even though the generated item bundle effects are still present, their impact is not that of LID but that of added independent error variances on the items' answers. This is in contrast to the item bundle effects in the sequential test designs in which the items of an item bundle commonly influence the same dimension and, therefore, not only result in added random error variance but in common error variance (that is, they introduce LID into the data).

The parallel test design's characteristic that it does not include LID in its actual sense is emphasized by the results depicted in Figure 2. Here, the multidimensional calibration of the data for the parallel test design for correlations of .9 and large item bundle effects results in a worse model fit than the calibration of the unidimensional model. A result

which is theoretically impossible if the test data responds to the assumptions of IRT. It is attributable to the fact that the unidimensional calibration includes LID and therefore overestimates its model fit while the multidimensional calibration (for the parallel test design) does not include the LID and therefore does not overestimate its fit. For the same reason the multidimensional calibration always provides a better model fit for the sequential test design than for the parallel test design.

The origin for the biases in the estimation of the correlations is demonstrated via the differing impacts on the variance components depicted in Figure 4. In all cases the LID included in the calibration of the data for the sequential test design results in an increase in the unidimensional and the dimension-specific variance component in comparison to the parallel test design. The increase of the two variance components is different, however. Since the single dimensions only include 20 items, the four items of an item bundle have a comparatively large dimension-specific effect; while the effect on the unidimensional variance component that is determined by a total of 80 items is comparatively smaller. The unidimensional variance components for calibrations of the sequential and the parallel test designs, therefore, only differ to a small amount, while the subtest-specific variances components show substantial differences. Particularly the results for correlations of .9, in which the subdimension-specific variances for the sequential test design increase beyond the actually existing[4], hereby, emphasize that variance is introduced into the measures that is solely due to the used item administration form and not due to the variance of the measured constructs. The presence of LID might therefore result in a false interpretation of differences between measured dimensions; assuming that differences are solely attributable to differences in the constructs while they might, in fact, partially origin in item bundle effects. The results in Figure 4 show that for a sequential test design with correlations of .7 and .9 already medium size item bundle effects result in dimension-specific variances that originate only by 69.7% and 56.5%, respectively, in the measured construct (0.30 of 0.43 and 0.13 of 0.23).

Conclusion

The results of the presented simulation study emphasize that LID not only biases the results of unidimensional calibrations but additionally biases the covariance estimates in multidimensional calibrations. Moreover, the chosen test design for the measurement of the multidimensional construct interferes with the impact of the LID and defines the direction of the bias. Considering that in practice test designs commonly are not as strict as the designs presented here but might consist of a mixture of item bundles that are attributed sequentially or parallel to different dimensions, the effect of the LID will often

[4] The reliabilities of the subdimension-specific components are very low for correlations of .9. The added random variances due to the LID, therefore, only have a small impact here in contrast to the added common variances. For correlations of .7 and .5 the reliabilities of the subdimension-specific variance components increase; therefore, the added random variances there have a relatively larger impact, which prevents increases in the variances as observed for correlations of .9.

be hard to predict. The differences in the results for the two presented test designs, however, show that the effects due to local item dependencies have to be separated into two different types of effects: (1) their effect as an error variance (visible via the results of the parallel test design) and (2) their effect as a redundantly modeled part of the measured latent trait (visible via the difference between the results for the parallel and the sequential test design).

Furthermore, the results underline the importance of the investigation of LID during test construction and test analysis in order to prevent an interpretation of biased results.

References

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

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

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

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments* (Vol. 1, pp. 51-70). Princeton, NJ: IEA-ETS Research Institute.

Brandt, S. (2010). Estimating tests including subtests. *Journal of Applied Measurement, 11*, 352-367.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*, 41-54.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*, 3-21.

Marsh, H. W., & O'Neill, R. (1984). Self Description Questionnaire III: the construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement, 21*, 153-174.

Martin, M. O., Gregory, K. D., & Stemler, S. E. (Eds.). (2000). *TIMMS 1999 technical report*. Chestnut Hill, MA: Boston College.

Maul, A. E. (in press). Examining the structure of emotional intelligence at the item level: New perspectives, new conclusions. *Cognition and Emotion*.

OECD. (2002). *PISA 2000 technical report*. Paris: OECD.

OECD. (2005). *PISA 2003 technical report*. Paris: OECD.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4, pp. 321-333). Berkeley: University oc California Press.

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

Rosenbaum, P. R. (1988). Items Bundles. *Psychometrika, 53*, 349-359.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion*. Göttingen: Verlag Hans Huber.

Rost, J., & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement, 26*, 42-56.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

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

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*, 247-260.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

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

Wang, W.-C., Cheng, Y.-Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement, 65*, 5-27.

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

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

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). ConQuest: Generalized item response modelling software. Camberwell, Victoria: Australian Council for Educational Research.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. M. (1993). Scaling performance assessments – strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Appendix A

-2 log likelihoods for the four-dimensional constructs with correlations of .5 (Table A1), .7 (Table A2), and .9 (Table A3).

Table A1

	No bundle effect	Bundle effect with variance of 0.5	Bundle effect with variance of 1.0	Bundle effect with variance of 2.0
Unidimensional				
Sequential	85299.7	86384.0	89134.0	95633.7
Parallel	85299.7	86386.0	89153.5	95673.0
Multidimensional				
Sequential	80921.7	82081.1	85039.0	91878.1
Parallel	80921.7	82408.4	86134.9	94507.4
Subdimensional				
Sequential	80927.4	82088.4	85047.5	91880.8
Parallel	80927.4	82416.2	86142.6	94514.5

Table A2

	No bundle effect	Bundle effect with variance of 0.5	Bundle effect with variance of 1.0	Bundle effect with variance of 2.0
Unidimensional				
Sequential	82229.5	83483.7	86653.0	94060.7
Parallel	82229.5	83472.2	86666.6	94080.5
Multidimensional				
Sequential	80055.4	81272.9	84360.9	91483.2
Parallel	80055.4	81542.6	85312.0	93721.5
Subdimensional				
Sequential	80063.7	81281.3	84367.4	91489.6
Parallel	80063.7	81550.9	85319.5	93736.9

Table A3

	No bundle effect	Bundle effect with variance of 0.5	Bundle effect with variance of 1.0	Bundle effect with variance of 2.0
Unidimensional				
Sequential	78931.9	80410.4	84040.2	92406.3
Parallel	78931.9	80378.6	84033.2	92418.8
Multidimensional				
Sequential	78542.3	79926.4	83301.0	90921.3
Parallel	78542.3	80067.9	83891.7	92559.5
Subdimensional				
Sequential	78567.2	79944.9	83313.7	90930.2
Parallel	78567.2	80095.1	83928.1	92619.6

Appendix B

Estimated correlations using the multidimensional model

	No bundle effect	Bundle effect with variance of 0.5	Bundle effect with variance of 1.0	Bundle effect with variance of 2.0
Sequential				
Correlation of .5	0.51	0.50	0.47	0.38
Correlation of .7	0.71	0.69	0.65	0.53
Correlation of .9	0.89	0.87	0.83	0.67
Parallel				
Correlation of .5	0.51	0.51	0.53	0.59
Correlation of .7	0.71	0.71	0.72	0.77
Correlation of .9	0.89	0.89	0.90	0.90

Appendix C

Estimated variances for the generated simulation data

	No bundle effect	Bundle effect with variance of 0.5	Bundle effect with variance of 1.0	Bundle effect with variance of 2.0
Sequential				
Correlation .5				
Dimension (MD)	1.98	1.87	1.63	1.18
Subdimension (SD)	0.73	0.70	0.65	0.55
Main Dimension (SD)	1.25	1.17	0.98	0.63
Correlation .7				
Dimension (MD)	1.99	1.89	1.64	1.19
Subdimension (SD)	0.44	0.44	0.43	0.42
Main Dimension (SD)	1.56	1.45	1.22	0.77
Correlation .9				
Dimension (MD)	2.02	1.93	1.67	1.21
Subdimension (SD)	0.19	0.20	0.23	0.30
Main Dimension (SD)	1.83	1.76	1.46	0.91
Parallel				
Correlation .5				
Dimension (MD)	1.98	1.81	1.44	0.81
Subdimension (SD)	0.73	0.66	0.51	0.25
Main Dimension (SD)	1.25	1.15	0.94	0.56
Correlation .7				
Dimension (MD)	1.99	1.83	1.46	0.83
Subdimension (SD)	0.44	0.40	0.30	0.15
Main Dimension (SD)	1.56	1.44	1.16	0.69
Correlation .9				
Dimension (MD)	2.02	1.88	1.52	0.90
Subdimension (SD)	0.19	0.17	0.13	0.08
Main Dimension (SD)	1.83	1.75	1.42	0.84

Note. MD = Result of the multidimensional model; SD = Result of the subdimension model.