Psychological Test and Assessment Modeling, Volume 52, 2010 (3), 323-340

Modeling the multidimensional structure of students' foreign language competence within and between classrooms

Jana Höhler[1], *Johannes Hartig*[2] & *Frank Goldhammer*[2]

Abstract

Combining multilevel (ML) analysis and multidimensional item response theory (MIRT) provides a valuable method for analyzing data of educational assessments, where clustered data (e.g., students in classes) and multidimensional constructs frequently occur. It allows to model multiple ability dimensions while simultaneously taking the hierarchical structure into account. The dimensional structure of students' foreign language competence within and between classrooms was investigated by applying a ML-MIRT measurement model to data of $N = 9{,}410$ students in 427 classes who had answered three different subtests of English as a foreign language. Results were compared to a MIRT model not taking into account the multilevel structure. A markedly more differentiated correlation structure is found within classrooms compared with the between-classroom level and compared with the model without multilevel structure. Results show that by modeling the latent multilevel structure, estimation and interpretation of ability profiles can be possible even with highly correlated ability dimensions.

Key words: item response theory, multidimensional item response theory, multilevel analysis, models of competencies, English as a foreign language

[1] *Correspondence concerning this article should be addressed to:* Jana Höhler, PhD, Center for Research on Educational Quality and Evaluation, German Institute for International Educational Research (DIPF), Schloßstraße 29, 60486 Frankfurt, Germany; email: hoehler@dipf.de

[2] German Institute for International Educational Research, Frankfurt, Germany

Introduction

The application of Item Response Theory (IRT) for modeling competencies in psychological and educational contexts is well established (Embretson & Reise, 2000; Lord & Novick, 1968; van der Linden & Hambleton, 1997; Wilson, 2005). If the focus lies on a detailed assessment of a relatively broad competence (e.g., foreign language competence) a common strategy is to break it down into narrower unidimensional abilities (e.g., reading and listening comprehension in a foreign language). These multiple components can then be measured using separate unidimensional IRT models. In this approach, the similarities and differences of these specific abilities are neglected, or at least, not explicitly modeled. Nevertheless, the interpretation of such separate unidimensional models is straightforward, and thus they may be advantageous especially if the objective of the assessment is to describe levels of performance in a relatively narrow ability like reading comprehension (Hartig & Höhler, 2008).

Another approach is to apply multidimensional IRT (MIRT) models, simultaneously incorporating multiple latent ability dimensions representing the broader competence, a procedure with an “elegant simplicity” (Reckase, 1997b, p. 25). Theoretically the multidimensional approach is more accurate than the unidimensional one, because in reality there are always nonzero correlations between latent traits (Cheng, Wang, & Ho, 2008), particularly regarding specific cognitive abilities. Moreover, measurement precision increases with the number of dimensions and their intercorrelations (Wang, Chen, & Cheng, 2004; Yao & Boughton, 2007). The predominantly questionable assumption of unidimensionality is not addressed within this paper, but one should be aware that, as Reckase (1997a) states, “the number of dimensions is often underestimated and that overestimating the number of dimensions does little harm” (p. 274).

Another issue for modeling competencies within educational contexts is that the assessed data are often structured hierarchically, meaning that the sample consists for example of students who are clustered in classes. Because sampling schools and/or classrooms for educational studies is often more feasible and economic than sampling individual students, the frequency of cluster randomized trials in education is likely to increase in the future (Spybrook, 2008). Multilevel modeling provides an adequate methodology for analyzing such hierarchically structured data (e.g., Hox, 2002; Kreft & de Leeuw, 1998).

Combining the (M)IRT and multilevel modeling approach allows accommodating the dependency typically found in clustered data. Furthermore, it enables measurement of (a) latent traits at different levels, (b) the (co-)variance decomposition of the latent traits at different levels, and (c) the estimation of relationships between predictor variables and latent traits at different levels (Pastor, 2003).

In the following we firstly provide a brief introduction into 1) MIRT, 2) multilevel modeling, and 3) the combination of these two approaches. Then we provide an empirical example for modeling foreign language competencies applying this combined multilevel MIRT (ML-MIRT) approach. Finally, results of the empirical application and implications of ML-MIRT modeling for interpreting and reporting test scores are discussed.

Multidimensional Item Response Theory

In IRT item responses are modeled as a function of individual trait levels and item properties (e.g., difficulty). Both, individual trait levels and item difficulties, can be described herein on a common scale. All IRT models used in this paper are logistic models, with the logit function defined as

$$\text{logit}(\mu) \equiv \frac{e^{\mu}}{1+e^{\mu}}. \tag{1}$$

In the one-parameter logistic (1PL) IRT model or Rasch model (Rasch, 1960) for dichotomous responses, the probability of a correct response of person p on item i ($x_{pi} = 1$) is modeled as a function of the individual ability θ_p of person p and the item difficulty b_i of item i:

$$\text{P}\left(x_{pi} = 1 \middle| \theta_p, b_i\right) = \text{logit}\left(\theta_p - b_i\right). \tag{2}$$

In MIRT, the probability of a correct response does not depend on a single ability variable θ, but on a vector $\boldsymbol{\theta}$ of K multiple latent ability dimensions θ_k. The multidimensional generalization of the 1PL model can be written as

$$\text{P}\left(x_{pi} = 1 \middle| \boldsymbol{\lambda}_i, \boldsymbol{\theta}_p, b_i\right) = \text{logit}\left(\boldsymbol{\lambda}_i' \boldsymbol{\theta}_p - b_i\right). \tag{3}$$

Here, $\boldsymbol{\lambda}_i$ is a $K \times 1$ vector of fixed factor loadings with $\lambda_{ik} \in \{0,1\}$, defining the influence of the K different abilities on the probability to solve item i, and b_i again is the difficulty of item i. $\boldsymbol{\lambda}_i' \boldsymbol{\theta}_p$ is the sum of the latent abilities relevant for item i.

Empirical analysis of a MIRT model provides latent estimations for the variance $\sigma^2_{\theta_K}$ for each dimension θ_k as well as the covariance structure $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ of the ability dimensions.

Multilevel modeling within educational assessments

In educational assessments data are often structured hierarchically, meaning that the sample consists for example of students (within-cluster level, L1) who are clustered in classes (between-cluster level, L2). Here, subjects are not sampled individually and randomly from the population of interest. However, this supposition underlies most statistical analysis approaches (e.g., regression analysis). Violating the assumption of independent observations may be especially crucial for educational achievement tests, because students from the same classroom are likely to share strong common sources of variation (Muthén, 1991). Analyzing clustered data without considering the hierarchical structure leads to various problems which are comprehensively described in numerous textbooks (e.g., Hox, 2002; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002; Skrondal & Rabe-Hesketh, 2004; Snijders & Bosker, 1999). Such a complex sample design can be regarded as complication to the statistical analysis. However, it can also provide an opportunity for more informative modeling of substantive phenomena, like exploring relationships among variables located at different levels simultaneously (Muthén, 1991).

Adequate methodology for the analysis of hierarchically structured data is particularly important for research questions focusing on persons embedded within a social system (like students in classrooms). Such research questions occur frequently within educational studies. Specific factors influencing the individual can be assumed to be potentially significant on each level of such a hierarchical system. For instance, the teaching style of the teacher is a factor at the between-cluster level (L2 predictor variable), while variables related to the students' family background are factors at the within-cluster level (L1 predictor variables). The individual performance can thus be regarded as affected by class membership itself (usually modeled as random effect) as well as by L1 and L2 predictor variables (fixed effects).

In multilevel modeling, an observed variable y_{pg} for person p in group g is decomposed in the sum of the grand mean, the variation of the group mean from the grand mean (between-cluster variation y_g^{B}) and the variation of the individual values of y from the group mean (within-cluster variation y_{pg}^{W}):

$$y_{pg} = \overline{y}_{..} + y_g^{\mathrm{B}} + y_{pg}^{\mathrm{W}}\,, \tag{4}$$

$$y_g^{\mathrm{B}} = \overline{y}_{.g} - \overline{y}_{..}\,, \tag{5}$$

$$y_{pg}^{\mathrm{W}} = y_{pg} - \overline{y}_{.g}\,. \tag{6}$$

Here $\overline{y}_{.g}$ is the mean of group g and $\overline{y}_{..}$ is the mean of y across the whole sample (grand mean). The variance of y is decomposed into variance between groups and variance within groups:

$$\sigma_y^2 = \sigma_{y^{\mathrm{B}}}^2 + \sigma_{y^{\mathrm{W}}}^2\,. \tag{7}$$

To gain a measure of how similar persons within the same group are, or in other words how strong the effect of group membership is on y, the intraclass correlation coefficient (*ICC*) can be calculated. The *ICC* is a coefficient which is of interest when effects within social systems are analyzed, for example in research on students within classrooms. It is defined as the relative size of the between-cluster variance to the total variance of y:

$$ICC_y = \frac{\sigma_{y^{\mathrm{B}}}^2}{\sigma_{y^{\mathrm{B}}}^2 + \sigma_{y^{\mathrm{W}}}^2}\,. \tag{8}$$

If there is clustering to a certain degree, even if it may seem very small (e.g., *ICC* = .01, or *ICC* = .05), the actual alpha level of statistical default tests will increase dramatically, and the more the *ICC* and the sample size increase the more the alpha inflation will increase too. This is because the assumption of independent observations is violated, which results in negatively biased standard errors. These underestimated standard errors lead to overestimation of significance or alpha inflation (Cohen, Cohen, West, & Aiken, 2003). Thus, considering the hierarchically structured data is at least important to avoid methodological artifacts and misinterpretations. This can be achieved by adjusting standard errors with appropriate estimation algorithms that take the *ICC*s of the analyzed variables into account (e.g., Snijders & Bosker, 1999). To additionally gain differentiated information for the different hierarchical levels of the data structure, multilevel analysis can be applied. Here, the hierarchical structure is explicitly modeled and variables measured on different levels of the hierarchy can be simultaneously included in the analysis.

For the multivariate case not only the variance of a single variable, but the whole variance-covariance matrix of multiple observed variables is decomposed. In multivariate multilevel modeling and multilevel structural equation modeling (ML-SEM), respectively, a between-cluster covariance matrix $\mathbf{\Sigma}^{\mathrm{B}}$ and a within-cluster covariance matrix $\mathbf{\Sigma}^{\mathrm{W}}$ can be computed. These covariance matrices have the property that they are orthogonal and additive (cf. Hox, 2002):

$$\mathbf{\Sigma}_y = \mathbf{\Sigma}_y^{\mathbf{B}} + \mathbf{\Sigma}_y^{\mathbf{W}} . \tag{9}$$

Mehta and Neale (2005), for instance, provide a didactical explanation of ML-SEM and demonstrate the equivalence of this approach and general mixed-effects models. In Hox (2002) a discussion of advantages and disadvantages to the multivariate multilevel approach is given.

Combining MIRT and multilevel models

Typically, the hierarchical structure in clustered datasets is not considered by traditional measurement models, like classical test theory or IRT (Kamata, Bauer, & Miyazaki, 2008). Combining MIRT and multilevel models for the psychometric analysis provides several advantages. First and most important, clustered data are analyzed appropriately by taking into account both within- and between-cluster variations. For example, Raudenbush, Rowan, and Kang (1991) showed that the well-known and established Cronbach's alpha coefficient is inherently ambiguous if clustering in educational studies is ignored, because it measures neither the reliability of L1 measures nor the reliability of L2 measures (see also Kamata et al., 2008). Second, this combined approach offers the opportunity to incorporate covariates and interaction effects (Kamata et al., 2008; Muthén, 1991). Examples for embedding Rasch based multilevel IRT in general approaches are given by Kamata (2001) as well as by Kamata et al. (2008) for the hierarchical generalized linear model (HGLM). Limitations of these approaches concern a) the assumption of equal or *a priori* known item discriminations for all test components, and b) the simultaneous modeling of several latent variables, where the covariances typically would be left unstructured, meaning that each dimension is correlated with every other dimension and that there are no structural relations between the dimensions, in contrary to ML-SEM approaches. Kamata and Cheong (2007) generalize the HGLM approach for multidimensional constructs within the generalized linear mixed model and demonstrate with an empirical application how to study the relationships between covariates at different levels and the constructs of interests.

Skrondal and Rabe-Hesketh (2004) provide a formulation and application of a four-level two-dimensional item response model with a logit link for dichotomous items in the generalized random coefficient notation (Generalized, Linear, Latent and Mixed Model; GLLAMM) for attitudes to abortion. Fox (2005a, b; Fox & Glas, 2001) provides an in-depth analysis of estimation algorithms for multilevel IRT models focusing on the fully Bayesian procedure. Fox (2004; see also Fox & Glas, 2001; 2003) defines a multilevel IRT model with a latent dependent variable measured by an IRT model. Here, item responses are regarded as L1 units, and item responses are nested within the students. Accordingly, a three level model results with item responses on L1, students on L2, and groups (e.g., classes or schools) on level 3.

In the procedure applied here we define a two-level measurement model for MIRT, with L1 as the 'student'-level and L2 as the 'group'-level (i.e., classes). In this model the latent ability θ_{pgk} of a person p in group g for dimension k is decomposed in the sum of the grand mean ($\overline{\theta}_{..k}$), in the variation (θ^{B}_{gk}) of the group mean of dimension k from the respective grand mean $\overline{\theta}_{..k}$, and the variation of the individual ability from the group mean (within-variation $\theta^{\mathrm{W}}_{pgk}$ of dimension k):

$$\theta_{pgk} = \overline{\theta}_{..k} + \theta^{\mathrm{B}}_{gk} + \theta^{\mathrm{W}}_{pgk} , \tag{10}$$

$$\theta^{\mathrm{B}}_{gk} = \overline{\theta}_{.gk} - \overline{\theta}_{..k} , \tag{11}$$

$$\theta^{\mathrm{W}}_{pgk} = \theta_{pgk} - \overline{\theta}_{.gk} . \tag{12}$$

Note that the grand mean $\overline{\theta}_{..k}$ is typically restricted to zero for identification purposes.

The probability of a correct answer of person p in group g on item i ($x_{pgi} = 1$) depends on the weighted sum of the ability components (subscripts in the conditional part of this and the following formulas are omitted for convenience):

$$P\left(x_{pgi} = 1 \middle| \boldsymbol{\lambda}, \boldsymbol{\theta}^{\mathrm{B}}, \boldsymbol{\theta}^{\mathrm{W}}, b\right) = \operatorname{logit}\left(\boldsymbol{\lambda}_i' \left(\boldsymbol{\theta}^{\mathrm{B}}_g + \boldsymbol{\theta}^{\mathrm{W}}_{pg}\right) - b_i\right). \tag{13}$$

Here $\boldsymbol{\lambda}_i$ is a $K \times 1$ vector of fixed factor loadings defining the influence of the K different abilities on item i for both levels. $\boldsymbol{\theta}^{\mathrm{B}}_g$ is a vector of deviations of group means from the grand mean for group g in all K ability dimensions, and $\boldsymbol{\theta}^{\mathrm{W}}_{pg}$ is the vector of L1 deviations of person p from the group means in all K ability dimensions. b_i again is the difficulty of item i.

The model in Equation (13) assumes identical dimensional structure and loading pattern on L1 and L2. However, theoretically the number of dimensions as well as the loading patterns can also vary between between-cluster level and within-cluster level, leading to a more general model:

$$P\left(x_{pgi}=1\middle|\boldsymbol{\lambda}^{\mathrm{B}},\boldsymbol{\lambda}^{\mathrm{W}},\boldsymbol{\theta}^{\mathrm{B}},\boldsymbol{\theta}^{\mathrm{W}},b\right)=\operatorname{logit}\left(\boldsymbol{\lambda}_i^{\mathrm{B}\prime}\boldsymbol{\theta}_g^{\mathrm{B}}+\boldsymbol{\lambda}_i^{\mathrm{W}\prime}\boldsymbol{\theta}_{pg}^{\mathrm{W}}-b_i\right). \tag{14}$$

$\boldsymbol{\lambda}_i^{\mathrm{B}}$ is here a $K \times 1$ vector of fixed factor loadings defining the influence of the K different abilities on item i on group level, and $\boldsymbol{\lambda}_i^{\mathrm{W}}$ is a $K \times 1$ vector of fixed factor loadings defining the influence of the K different abilities on item i within groups. In such cases the decomposition of $\boldsymbol{\theta}_p$ as shown before is no longer possible, because the latent dimensions represent different constructs and are measured by different items, respectively. For most models analyzed in this paper, we will assume that the number of dimensions K is identical for both levels and that the loading structure for all items also is identical for both levels (i.e., $K^{\mathrm{B}} = K^{\mathrm{W}}$, and $\boldsymbol{\lambda}_i^{\mathrm{B}} = \boldsymbol{\lambda}_i^{\mathrm{W}}$).

In ML-MIRT based on Equation (13), variances as well as the covariances of the latent dimensions can be decomposed in L1 and L2 proportions. The variance $\sigma^2_{\theta_k}$ of each ability dimension k is decomposed in within-cluster variance $\sigma^2_{\theta_k^{\mathrm{W}}}$ and between-cluster variance $\sigma^2_{\theta_k^{\mathrm{B}}}$:

$$\sigma^2_{\theta_k} = \sigma^2_{\theta_k^{\mathrm{W}}} + \sigma^2_{\theta_k^{\mathrm{B}}}. \tag{15}$$

The latent intraclass correlation coefficient $LICC_k$ for dimension k is defined as the proportion of between-cluster variance $\sigma^2_{\theta_k^{\mathrm{B}}}$ to the total variance:

$$LICC_k = \frac{\sigma^2_{\theta_k^{\mathrm{B}}}}{\sigma^2_{\theta_k^{\mathrm{W}}} + \sigma^2_{\theta_k^{\mathrm{B}}}}. \tag{16}$$

In the following we will use the abbreviation *ICC* for the 'manifest' intraclass correlation and *LICC* for the 'latent' intraclass correlation coefficient. The covariances of the latent abilities $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ are decomposed in between-cluster covariance and within-cluster covariance proportions:

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{\mathbf{B}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{\mathbf{W}}. \tag{17}$$

Empirical analysis of a ML-MIRT model provides latent estimates for the variance components $\sigma^2_{\theta_k^{\mathrm{B}}}$ and $\sigma^2_{\theta_k^{\mathrm{W}}}$ for each dimension k as well as the covariance structures $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{\mathbf{B}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{\mathbf{W}}$ on L1 and L2, respectively.

For research questions in educational contexts, where students' performance is assessed in hierarchical structured samples and modeled using an IRT-approach, combining these methods can provide valuable information (Fox, 2004; Fox, 2005a, b; Fox & Glas, 2001; Kamata et al., 2008). Using such a combined method allows, for instance, investigating how much variation of an ability dimension is determined by class membership. This decomposition provides a basis for calculating the *LICC* as the ratio of between-cluster variance to total variance. Furthermore, within a multidimensional model the decomposition of covariance between the dimensions for the different levels can be inspected. The proportion of covariance on L1 can be interpreted as the relationship between the deviation values of the students from the group mean for the respective dimensions or, in other words, the relations between the dimensions controlling for group membership. The L2 covariance proportion represents the relationship between the group means in the different dimensions.

Research aims

In this study, the latent correlation structure and (co-)variance decomposition of tests for English as a foreign language are examined for a representative German sample of ninth-graders. The first research aim is to demonstrate that the *LICC*s result in a higher estimate than the *ICC*s because measurement error is taken into account. Reflecting disattenuation for measurement error, the *LICC*s should exceed the *ICC*s for the specific ability dimensions, and the difference should decrease with increasing reliability of the applied test (Kamata et al., 2008; Lüdtke, Robitzsch, Asparouhov, Marsh, Trautwein, & Muthén, 2008; Raudenbush et al., 1991). Therefore the different reliabilities, *ICC*s, and *LICC*s for the three dimensions (reading comprehension, listening comprehension, and language awareness) are calculated. Different *LICC*s for the dimensions imply that class membership affects these specific abilities to a different extent, and thus indirectly confirm the assumption of substantive differences between the multiple dimensions in the model. For example, characteristics of class instruction could have a different impact on the *LICC*s, meaning that such characteristics influence the specific abilities to a different degree, and thus indicating a conceptually different meaning of the latent variables. A further interesting point is how much variance of the ability dimensions is determined by class membership. For all dimensions, a high amount of variance explained by class membership (between-cluster level) is expected (*LICC* > .50), because of the high selectivity of schools based on early tracking in the German school system. High *ICC*s are frequently found for the German school system (e.g., Organisation for Economic Co-operation and Development [OECD], 2003, 2007). In PISA 2006, for example, an *ICC* of even .80 was found for students' performance in the reading scale (OECD, 2007[3]).

Another research interest concerns the correlations between the three dimensions on the different levels. In general, there is a strong tendency to unidimensionality of language assessment data (Carroll, 1993; Diakidoy, Stylianou, Kerefillidou, & Papageorgiou, 2005; Jude et al., 2008). For the German school system the general language competence level is strongly dependent on class membership, because students are tracked very early (around the age of nine) among three different school types. We expect a high amount of covariance between the three ability dimensions on between-cluster level, because classes show a relatively homogenous performance across these specific abilities. For example, classes with a high competence in one of the dimensions have also a high competence in the other dimensions. As class membership is controlled within the ML-MIRT approach it is expected that a more differentiated correlation structure between the spe-

[3] cf. Table 4.1d, available at: http://www.pisa.oecd.org/dataoecd/30/62/39704344.xls; retrieved on 10-19-2009.

cific abilities can be found on L1 compared with L2 as well as compared with a MIRT approach without modeling the hierarchical data structure.

Method

Data and testing constructs

The data for the empirical application come from a large-scale assessment of 9th-grade students' language competencies named DESI[4]. In DESI, a number of tests were developed to assess different dimensions of students' language competencies in German and English as a foreign language (EFL), primarily in Germany, but also in Austria and the predominantly German-speaking North-Italian province of Bolzano-Bozen. The analysis is based on the German data from the tests of listening comprehension, reading comprehension, and language awareness (grammar) in EFL. The tests of language awareness, reading and listening comprehension were chosen because they constitute a differentiated and comprehensive picture of receptive foreign language competence.

The English listening comprehension test (Nold & Rossa, 2007a) comprises 51 multiple-choice questions with three or four response categories referring to six texts and dialogues presented on audio tape. For each text or dialogue, between one and ten items should be answered. The questions ask for the comprehension of gist and details of the texts. The listening comprehension construct in EFL focuses on the real-time processing of spoken scripts.

The aim of the questions for English reading comprehension (Nold & Rossa, 2007b) is to assess the comprehension of gist and details of texts presented in written English. The test contains 46 multiple-choice questions with four response categories, referring to four written texts. For each of the texts, ten or twelve questions are given.

Language awareness (grammar; Nold & Rossa, 2007c) is somewhat more distinct from the former two testing constructs. Here, the ability to complete sentences with the aid of given alternatives and to identify incorrect grammatical structures is assessed. Overall, the testing construct language awareness (grammar) for EFL is orientated on the ability to correct oneself in a formal and communicative linguistic manner. The test provides 29 item scores based on two item formats. There are usual multiple-choice items with four possible responses, of which one is correct. Additionally, there are tasks demanding a decision, whether there is a mistake in one of three defined expressions. Items in all three tests were scored as either correctly or incorrectly answered, thus making an IRT model for dichotomous responses adequate for the analysis.

The items for listening and reading comprehension were presented in a matrix design (e.g., Frey, Hartig, & Rupp, 2009). On average, each student answered 15 of the 51 listening comprehension items, and 20 of the 46 reading comprehension items. Overall,

[4] *DESI* (Beck & Klieme, 2007; Klieme et al., 2008) is a German acronym for "German English Student Assessment International" [Deutsch Englisch Schülerleistung International].

$N = 9{,}410$ students within $M = 427$ classes with at least two valid responses in each of the three tests were included in the analysis. For all analyses sample weights to adjust for unequal selection probabilities were used.

Model description

The specified model contains three correlated dimensions on two levels, one for each testing construct (listening comprehension, reading comprehension, and language awareness). The first level is the within-cluster level for student responses (L1) and the second level is the between-cluster level (L2). A schematic illustration of the model is given in Figure 1.

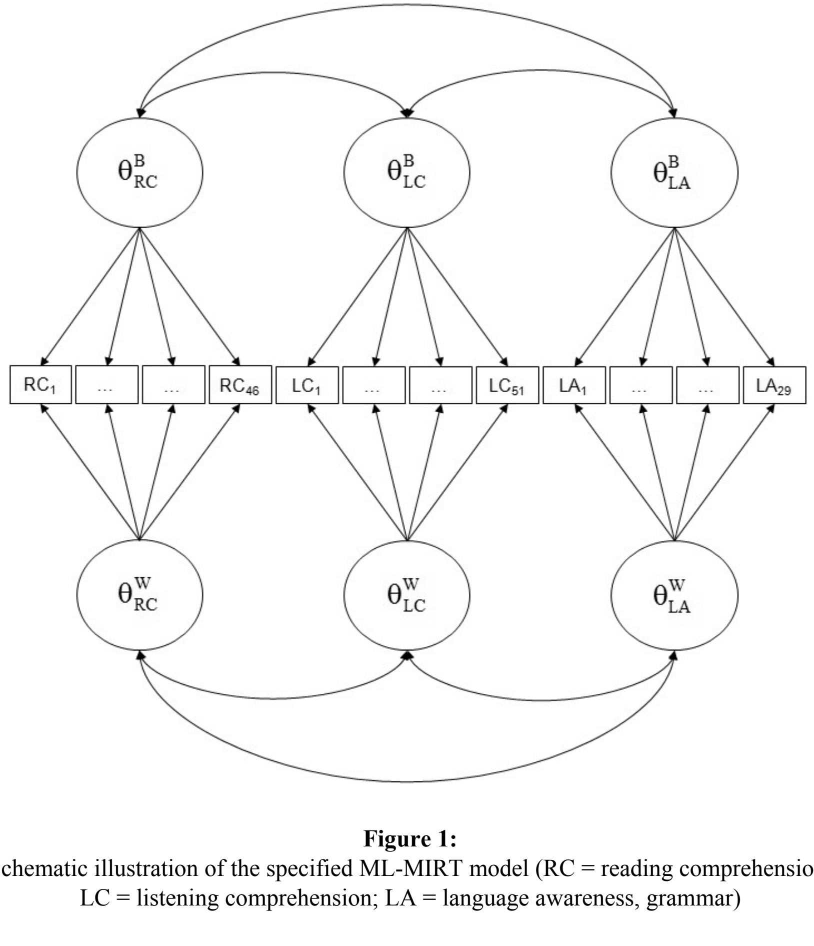

Figure 1:
Schematic illustration of the specified ML-MIRT model (RC = reading comprehension; LC = listening comprehension; LA = language awareness, grammar)

Additionally, a model with three correlated dimensions was calculated without explicitly considering the multilevel structure. Here, the *Mplus* sandwich estimator to adjust for underestimated standard errors was applied (Muthén & Muthén, 1998-2007b).

The *ICC*s were calculated on the base of weighted likelihood estimates (WLEs; Warm, 1989) obtained from unidimensional scaling of each test with *ConQuest* (Wu, Adams, & Wilson, 1998). All (ML-)MIRT models were analyzed with *Mplus* Version 5.1 (Muthén & Muthén, 1998-2007a) using maximum likelihood estimation with robust standard errors and for the ML-MIRT models using Montecarlo-Integration with 1000 nodes per dimension. The *LICC*s were calculated according to Equation (16) on the base of the estimated variance components obtained from the ML-MIRT analysis.

Results

Latent intraclass correlations and variance decomposition

The first research aim is to demonstrate that the calculated *LICC*s for the different dimensions exceed the *ICC*s, reflecting disattenuation for measurement error. Thus, the difference between the *ICC*s and *LICC*s should decrease with increasing reliability of the applied test. As expected, results show that the *LICC*s exceed the respective *ICC*s for the three dimensions listening comprehension, reading comprehension, and language awareness (see Table 1).

Furthermore, this difference becomes more pronounced with increasing measurement error. For example, listening comprehension was assessed with a relatively low reliability and shows the largest difference between *LICC* and *ICC*. However, there is also a quite large discrepancy in the *L/ICC*-difference between reading comprehension and language awareness, although these dimensions do not differ much in their test reliability.

Table 1 shows also the variance components, by which the *LICC*s were calculated (see Equation 16). As expected, the calculated *LICC*s indicate that more than 50% of the variance is explained by class membership for each dimension. For listening comprehension the highest *LICC* results with 77% explained variance by class membership.

Table 1:
Variance components on the within- and between-cluster level and intraclass correlations for the three dimensions

	$\sigma^2_{\theta_K^{\mathrm{W}}}$	$\sigma^2_{\theta_K^{\mathrm{B}}}$	*LICC*	*ICC*	Reliability*
Reading Comprehension	0.51	1.05	.68	.52	.74
Listening Comprehension	0.26	0.86	.77	.54	.66
Language Awareness	0.51	1.29	.72	.66	.77

Notes: $\sigma^2_{\theta_k^{\mathrm{W}}}$ = within-cluster variance, $\sigma^2_{\theta_k^{\mathrm{B}}}$ = between-cluster variance, *ratio of average posterior variance to estimated population variance.

Correlations between the ability dimensions

For the correlations between the latent variables it is expected that controlling class membership within the ML-MIRT approach provides a more differentiated ability structure. Accordingly, the correlations in a MIRT model and the correlations on between-cluster level should exceed the respective correlations on L1 of the ML-MIRT model. Considering the two-level structure within this ML-MIRT approach results indeed in a different latent correlation pattern between the three testing constructs for L1 and L2. As expected, a more differentiated structure is found on the L1 compared with the between-cluster level (see Table 2).

Table 2:
Latent correlations (standard errors) between reading comprehension (RC), listening comprehension (LC), and language awareness (LA). Correlations for the between-cluster level are printed above, for the within-cluster level below the main diagonal

	RC	LC	LA
RC		.971 (.030)	.995 (.033)
LC	.773 (.079)		.964 (.020)
LA	.533 (.032)	.612 (.040)	

The difference between the latent correlations on the different levels for reading comprehension and language awareness is notably high. Since the language awareness tasks to some extent also require reading comprehension a high correlation would be expected; however, the degree of correlation on L1 clearly justifies a separate assessment and interpretation of these two constructs. Interestingly, the highest correlation on the within-cluster level is found between reading and listening comprehension. These two tests require relatively similar cognitive processes like the decoding and understanding of English, the ability to process and integrate the information retrieved, as well as the comprehension of the gist and details of the presented text.

As all intercorrelations on L2 are quite high, an alternative model with one common dimension on the between-cluster level was analyzed and compared regarding model fit (see Table 3). The four-dimensional model fits worse than the six-dimensional model (χ^2_{Diff} = 5,346.87[5]; $df_{\mathrm{Diff}} = 5$; $p \leq .001$), so the analyses are based on the model with three dimensions on each level.

[5] The χ^2_{Diff}-value is calculated according to the procedure suggested by Satorra and Bentler (1999; see also htttp://www.statmodel.com/chidiff.shtml) and takes into account the scaling correction factor. The scaling correction factors are c0 = 2.096 for the model with 133 free parameters and c1 = 2.03 for the model with 138 free parameters, respectively.

Table 3:
Number of free parameters and fit indices for the four- and six-dimensional model

	Free parameters	LL	AIC	BIC
1 Between-, 3 Within-Dimensions	133	-332,318	664,902	665,430
3 Between-, 3 Within-Dimensions	138	-331,585	663,445	663,993

Notes: LL = Log-likelihood; AIC = Akaike's information criterion; BIC = sample-size adjusted Bayesian information criterion.

In a MIRT-model with adjusted standard errors the correlations between the three dimensions are also quite high (see Table 4). The ranking of the correlations in this model is the same as for the correlations on the within-cluster level for the ML-MIRT model. The coefficients exceed the respective values for the within-cluster level, but fall below the correlations on L2. So the ML-MIRT approach reveals a more differentiated correlation structure for the within-cluster level in comparison with the between-cluster level as well as in comparison with the MIRT-approach not considering the two-level data structure.

Table 4:
Latent correlations (standard errors) between reading comprehension (RC), listening comprehension (LC), and language awareness (LA)

	RC	LC
LC	.903 (.019)	
LA	.835 (.015)	.859 (.023)

Discussion

Within ML-MIRT it is possible to gain *LICC*s which are corrected for measurement error and hence comparable among each other, even when the test reliability differs for the different testing constructs. This is particularly valuable for comparing the impact of class membership on different ability dimensions when a relatively broad competence is assessed via multiple specific abilities.

For our empirical application results show, as expected, that class membership has a huge impact on test performance, as it explains between 68% and 77% of the total variance. Different values result for the *LICC*s regarding the different dimensions. This finding indicates a conceptually distinct meaning of the three latent variables in the model and indirectly confirms the assumption of substantive differences between these specific abilities, because class membership influences them to a different degree. In such cases, the conceptual and psychometric separation of specific abilities seems reasonable, despite of given high empirical correlations. Hence, the variance decomposition of variables can provide another cue -besides the empirical correlations- when deciding about the dimensional structure of a competence construct assessed with multilevel data. A further inspection of the variance components shows different values for all three dimensions on between-cluster level, but not for reading comprehension and language awareness on L1. However, the inverse conclusion that similar variance components confirm non-separability of dimensions cannot be drawn.

Within a multidimensional model, the decomposition of covariance between the dimensions for the different levels can also provide valuable information. The correlations on within-cluster level can be interpreted as the relations between the specific ability dimensions reading comprehension, listening comprehension, and language awareness controlling for class membership. The between-cluster covariance proportion represents the relationship between the class means in these three dimensions. Correlations on between-cluster level should not be taken as evidence to confirm or reject a structural competence model on an individual level. Therefore, the findings on within-cluster level or of initial non-multilevel (M)IRT scaling approaches should be scrutinized. However, the correlational structure on between-cluster level should be taken into account if test scores are interpreted and reported on an aggregated level (i.e., when comparing classes or schools). For our empirical application the resulting high correlations on between-cluster level signify a low reliability of ability profile information concerning differences between classes. In such a case only the level of competence in the broader domain (i.e., English as a foreign language) should be reported when comparing performance of whole classes or schools, although the multidimensional IRT scaling may be preferable nevertheless. On within-cluster level, on the other hand, a differentiated feedback of students' ability profiles, with their individual strengths and weaknesses in reading comprehension, listening comprehension, and language awareness for improving students performance in EFL could be given with a satisfying reliability.

Furthermore, for individual feedback the ML-MIRT model provides the opportunity to communicate ability profiles with the position of the student relative to the group mean as well as the position of the group mean in relation to the grand mean. For a differentiated individual feedback both options should be used. In addition, it is possible to depict the relative position of a specific student not only to his/her own group mean, but also to other group means. That could be interesting, for example, if a different school type or class for this student is considered.

The structural equivalence of constructs at different levels of aggregation should be tested to avoid (dis)aggregation fallacies and ensure a stable psychological meaning of the traits at the different levels; although in multilevel applications for students' performance data the similarity of meaning across aggregation levels can often be taken for granted (van de Vijver & Poortinga, 2002). According to Muthén (1991, 1994) a construct is equivalent across aggregation levels if a model that postulates the same number of factors at each level, and imposes the same relationship to all variables at both levels, shows a good fit. However, if the dimensional structure and/or loading pattern differs across levels, interpretation of the theoretical and conceptual meaning of the assessed construct for the different levels becomes quite complex.

To summarize, analyzing multidimensional clustered data with a ML-MIRT measurement model seems to be a quite promising procedure. First, it is possible to analyze multiple abilities simultaneously. Of course, this holds true for all MIRT models. But second, in comparison with a MIRT model, it has the advantage that L1 and/or L2 predictor variables could be directly included and investigated (see e.g., Fox, 2004; Fox & Glas, 2001; 2003; Kamata et al., 2008; Muthén, 1991). Third, compared with a manifest variable approach, it yields more accurate estimates, as demonstrated by Lüdtke et al. (2008) for contextual studies. Besides these more technical advantages, a ML-MIRT analysis enables a more differentiated individual feedback of ability profiles. Moreover, it allows to test the structural equivalence of a construct across different levels of aggregation and thus, to investigate how differentiated a feedback on a specific level should be given. For determining sufficient sample and cluster sizes and number of clusters to yield reliable estimates more simulation studies in this research field are required.

Author note

Jana Höhler, Center for Research on Educational Quality and Evaluation, German Institute for International Educational Research; Johannes Hartig, Center for Research on Educational Quality and Evaluation, German Institute for International Educational Research; Frank Goldhammer, Center for Research on Educational Quality and Evaluation, German Institute for International Educational Research.

Acknowledgments

The preparation of this paper was supported by grant HA 5050/2-1 from the German Research Foundation (DFG) in the Priority Program "Models of Competencies for Assessment of Individual Learning Outcomes and the Evaluation of Educational Processes" (SPP 1293).

References

Beck, B., & Klieme, E. (Eds.) (2007). *Sprachliche Kompetenzen. Konzepte und Messung* [Language competencies: concepts and measurement]. Weinheim: Beltz.

Carroll, J. B. (1993). *Human cognitive abilities: a survey of factor-analytic studies.* New York and Cambridge: Cambridge University Press.

Cheng, Y.-Y., Wang, W.-C., & Ho, Y.-H. (2008). Multidimensional Rasch analysis of a psychological test with multiple subtests. *Educational and Psychological Measurement OnlineFirst,* published on September 3, 2008 as doi:10.1177/0013164408323241.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression / correlation analysis for the behavioral sciences* (Chapter 14: Random coefficient regression and multilevel models, pp. 536-367). Mahwah, N.J.: Lawrence Erlbaum.

Diakidoy, I. A., Stylianou, P., Kerefillidou, C. & Papageorgiou, P. (2005). The relationship between listening and reading comprehension of different types of text at increasing grade levels. *Reading Psychology, 26* (1), 55-80.

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

Fox, J.-P. (2004). Applications of multilevel IRT modeling. *School Effectiveness and School Improvement, 15*, 261-280.

Fox, J.-P. (2005a). Multilevel IRT model assessment. In L. A. van der Ark, M. A. Croon & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 227-252). Mahwah, NJ: Erlbaum.

Fox, J.-P. (2005b). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, 58, 145-172.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika, 66* (2), 271-288.

Fox, J.-P., & Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika, 68* (2), 169-191.

Frey, A., Hartig, J., & Rupp, A. (2009). Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice, 28*, 39-53.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within- and between-item multidimensionality. *Journal of Psychology, 2*, 89-101.

Hox, J. J. (2002). *Multilevel Analysis: Techniques and applications*. Mahwah: Erlbaum.

Jude, N., Klieme, E., Eichler, W., Lehmann, R. H., Nold, G., Schröder, K., Thomé, G., & Willenberg, H. (2008). Strukturen sprachlicher Kompetenzen [Structures of language competencies]. In E. Klieme, W. Eichler, A. Helmke, R. H. Lehmann, G. Nold, H.-G. Rolff, K. Schröder, G. Thomé & H. Willenberg (Hrsg.). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* [Instruction and development of competence in German and English: results of the DESI study]. Weinheim: Beltz.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38* (1), 79-93.

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). Multilevel measurement modeling. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 345-388). Charlotte: Information Age Publishing.

Kamata, A., & Cheong, Y. F. (2007). Multilevel Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 217-232). New York : Springer.

Klieme, E., Eichler, W., Helmke, A., Lehmann, R. H., Nold, G., Rolff, H.-G., et al. (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* [Instruction and development of competence in German and English: results of the DESI study]. Weinheim: Beltz.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Welsley Publishing Company.

Lüdtke, O., Robitzsch, A., Asparouhov, T., Marsh, H. W., Trautwein, U., & Muthén, B. (2008). The multilevel latent covariate model: a new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13* (3), 203-229.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: multilevel structural equations modeling. *Psychological Methods, 10* (3), 259-284.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. Sociological Methods & Research, 22, 376-398.

Muthén, L. K., & Muthén, B. O. (1998-2007a). *Mplus statistical software.* Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (1998-2007b). *Mplus User's Guide* (4th Edition). Los Angeles, CA: Muthén & Muthén.

Nold, G., & Rossa, H. (2007a). Hörverstehen [Listening comprehension]. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung* [Language competencies: concepts and measurement] (pp. 178-196). Weinheim: Beltz.

Nold, G., & Rossa, H. (2007b). Leseverstehen [Reading comprehension]. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung* [Language competencies: concepts and measurement] (pp. 197-211). Weinheim: Beltz.

Nold, G., & Rossa, H. (2007c). Sprachbewusstheit [Language awareness]. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung* [Language competencies: concepts and measurement] (pp. 226-244). Weinheim: Beltz.

Organisation for Economic Co-operation and Development (2003). *The PISA 2003 assessment framework. Mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD.

Organisation for Economic Co-operation and Development (2007). *PISA 2006: Science Competencies for tomorrow's world (Volume 1: Analysis)*. Paris: OECD.

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: an illustration. *Applied Measurement in Education, 16* (3), 223-243.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (2nd Ed.). London: Sage Publications.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics, 16* (4), 295-330.

Reckase, M. D. (1997a). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern Item Response Theory* (pp. 271-286). NY: Springer.

Reckase, M. D. (1997b). The past and the future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25-36.

Satorra, A., & Bentler, P. M. (1999). *A scaled difference chi-square test statistic for moment structure analysis.* (UCLA Statistics Series No. 260). Retrieved from UCLA, Department of Statistics website: http://preprints.stat.ucla.edu/260/chisquare.pdf

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling. Multilevel, longitudinal and structural equation models.* Boca Raton: Chapman & Hall.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.

Spybrook, J. (2008). Power, sample size, and design. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 273-311). Charlotte: Information Age Publishing.

van de Vijver, F. J. R., & Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Psychology, 33*, 141-156.

van der Linden, W., & Hambleton, R. K. (1997). *Handbook of modern item-response theory*. Berlin: Springer.

Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Applied Psychological Measurement, 28*, 295-316.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Wilson, M. (2005). *Constructing measures. An item response modelling approach.* Mawah: Lawrence Erlbaum Associates.

Wu, M., Adams, R., & Wilson, M. (1998). *ConQuest: Generalized item response modelling software.* Melbourne: Australian Council for Educational Research.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*, 83-105.