Psychological Test and Assessment Modeling, Volume 52, 2010 (4), 354-379

Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis

Dominique P. Rauch[1] *& Johannes Hartig*[2]

Abstract

The dimensionality of a reading comprehension assessment with non-stem equivalent multiple-choice (MC) items and open-ended (OE) items was analyzed with German test data of 8523 9th-graders. We found that a two-dimensional IRT model with within-item multidimensionality, where MC and OE items load on a general latent dimension and OE items additionally load on a nested latent dimension, had a superior fit compared to an unidimensional model ($p \leq .05$). Correlations between general cognitive abilities, orthography and vocabulary and the general latent dimension were significantly higher than with the nested latent dimension ($p \leq .05$). Drawing back on experimental studies on the effect of item format on reading processes, we suppose that the general latent dimension measures abilities necessary to master basic reading processes and the nested latent dimension captures abilities necessary to master higher reading processes. Including gender, language spoken at home, and school track as predictors in latent regression models showed that the well known advantage of girls and mother-tongue students is found only for the nested latent dimension.

Key words: reading test, multiple-choice items, open-ended items, item response theory, components of reading ability

[1] *Correspondence concerning this article should be addressed to:* Dominique P. Rauch, research assistant, DIPF, German Institute of International Educational Research, Schloßstr. 29, 60486 Frankfurt am Main, Germany; email: rauch@dipf.de

[2] DIPF, German Institute of International Educational Research, Frankfurt am Main, Germany

Reading comprehension assessment with multiple-choice and open-ended items

For a long time reading has been a major area of interest for both practitioners and researchers in the fields of cognitive psychology, psycholinguistics, and educational assessment. While cognitive scientists are concerned with analysing the process of reading comprehension, educational researchers involved in assessment create measures for reading proficiency and try to explain different levels of reading proficiency, based on teacher, classroom, and student characteristics. A variety of reading tests have been developed, spanning basic and applied areas in psychology and education (Alderson, 2000). These tests comprise comprehension questions, recall and recognition tasks, and other kinds of dual-choice tasks, for example lexical decision (LD-) tasks; they furthermore contain on-line procedures such as eye-tracking and think aloud protocols (Zwaan & Singer, 2003).

To assess the readers' understanding of a text, educators and researchers use comprehension questions. This question type varies along several dimensions (Alderson et al., n.d.), among them the item format, which can be for example open-ended (OE) or multiple-choice (MC). In this article we bring together results from experimental psychological research on reading comprehension assessment and research on item formats in educational measurement. We strive to show that using a multidimensional IRT model can improve measurement in reading comprehension assessment and diagnostic usefulness of reading comprehension test scores. In the following sections we will first briefly outline the correspondence between psychological reading comprehension theory and educational measures of reading proficiency. This will be followed by a presentation of research on item formats and reading processes and corresponding text representations. In the third section we will summarize the main findings of two reviews and a meta-analysis on the effects of combining different item formats in educational assessment on the dimensionality of test performance. Finally we will argue that a multidimensional IRT model specifying a nested latent dimension for OE items has the potential to reflect the success in mastering higher-order reading processes.

Reading comprehension processes and reading proficiency measures

Reading assessment in educational contexts has been influenced by psychological reading comprehension theories, for example the construction-integration model (CI-Model, Kintsch, 1998; Kintsch & van Dijk, 1978). Kintsch describes comprehension as a sequence of cyclic, hierarchically nested processes, from the basics of letter- and word-recognition to higher-order processes of building a coherent representation of the meaning of a text. On the level of the cognitive representation of a text, the CI-model distinguishes between (1) a surface structure, which captures the exact wording of a text, (2) a text base, where propositions are connected, and (3) a situation model, which is a representation of the text, achieved through integration of text and prior knowledge. Accordingingly, measures of proficiency in reading assessment can be categorized with regard to the level of the comprehension process they address.

Measures of reading speed on the sentence level or cognitive tasks such as the lexical decision-task require processing at the word recognition level and the levels of the propositional text base (building local coherence). However, widely used reading tests, such as the Nelson-Denny reading test (Brown, Vick-Fishco, & Hannah, 1993), primarily target the propositional basis of the text as a whole, or the situation model. In fact, understanding a text in a way that allows the use of the information it contains in real-life activities requires an adequate situation model. Thus, large-scale assessments such as the *Program for International Student Assessment* (PISA; OECD, 2003) or *Progress in International Reading Literacy Study* (PIRLS; Mullis, Kennedy, Martin, & Sainsbury, 2006) aim to evaluate the quality of the construction of a propositional text-base and the situation model. The same holds true for DESI, a national assessment of language competencies in Germany (Beck & Klieme, 2007; DESI-Konsortium, 2008); The DESI reading proficiency test (Willenberg, 2007) aims to assess to what extent students are able to build a coherent representation of the text and to construct meaning by integrating previous knowledge.

In large scale assessments like PISA, PIRLS and DESI, reading proficiency measures are used to describe and compare the reading proficiency of groups of students. It is therefore essential that the used measures allow for valid inferences about reading proficiency. Moreover, as we will show in the following sections, more detailed measurement models, which distinguish between abilities needed to answer MC and OE item format, can possibly enrich this group-specific diagnostic information (Birenbaum & Tatsuoka, 1987). Some of the most prominent grouping variables, which are widely known to be associated with reading proficiency, are gender, language at home, and school track (specifically in school systems with early tracking, as in Germany). As it is stated in the executive summary of PISA 2006 results "reading is the area with the largest gender gaps" (OECD, 2007, p. 5). In large scale assessments girls usually obtain higher scores than boys in reading proficiency tests (OECD 2001, 2004, 2007). Test takers who speak a language other than the test language at home can be expected to perform lower in reading assessment than test takers who use the test language to communicate within their family (OECD, 2006; Stanat, Rauch, & Segeritz, 2010). The German school system segregates students after grade four, and large differences in reading test performance are found between the different tracks in secondary school (e.g. Artelt, Stanat, Schneider, & Schiefele, 2001).

Item formats and reading comprehension processes

Ever since standardized reading tests have been developed, item format has been an issue of discussion. MC items, which were widely accepted and thus were the dominant question format until the end of the 1980s, were then thought to be problematic for a number of reasons. Johns (1978), Katz, Lautenshalger, Blackburn, and Harris (1990), and later Daneman and Hannon (2001) pointed out that sometimes MC items can be answered without reading the respective text passage. This finding raised serious doubts about the validity of reading proficiency tests that are based solely on MC items. A further concern regarding MC items is that they make test takers select between pre-determined answers rather than allowing individualized responses as OE items do (Ozuru, Best, Bell, Witherspoon, & McNamara, 2007). What is more, OE items are considered to be closer to the reality of teaching and learning in the classroom, since they mirror teacher-student-communication on texts more closely.

In response to criticisms of MC items, test developers began to develop tests that combined different types of items, e.g. OE items and MC items. Today, this combination of item formats is found quite often in large scale assessments of reading comprehension and other subjects (Kim, Walker, & McHale, 2009; Rodriguez, 2003). In PISA 2000, 2003, and 2006, for example the proportions of open constructed response items, which required the test-takers to write down their answers, in all reading tasks were 45%, 50%, and 43%, while the proportion of simple MC items, for which test-takers chose one of several alternative answers, were 40%, 33%, and 29% (OECD, 2003, 2006). The rest were complex MC items and closed constructed response items. Including OE items into standardized reading proficiency tests opens the question of whether OE items measure the same construct as MC items, or rather different aspects of the reading comprehension process.

Shohamy (1984) examined the effect of item format (MC and OE items) and language of assessment (first language = L1 vs. second language = L2) on performance in second language (L2) reading tests. The main conclusion drawn from her study was that item format can affect test scores: MC items were found to be generally easier to answer than OE items. She attributed this to different language processes required to do the tasks.

Wolf (1993) examined the effects of different assessment tasks, languages of assessment and L2 language competence on L2 reading comprehension test performance. She compared effects of tests with multiple-choice items, open-answer items and cloze-tests. In cloze-tests test developers delete every nth word in as passage and test takers have to fill in each blank. Wolf found that the item type used to assess learners' reading comprehension affects their test results: test takers' performance on the multiple choice items was significantly better than that on the open-ended and cloze-test tasks.

Kobayashi (2002) examined the effects of text organization (association, description, causation, problem-solution) and task format (cloze-test, open-ended item, summary writing) on second language (L2) learners' performance in reading comprehension tests. Text organization and test format had a significant impact on test takers' performance. Additionally, Kobayashi found a statistically significant interaction between the two effects.

Other researchers (Cordon & Day, 1996; Pearson et al., 1999) investigated the impact of item format using think aloud procedures, i.e. asking participants to verbalize their thoughts while answering MC items and OE items. While Cordon and Day (1996) did not find any differences in cognitive processes underlying OE and MC items, Pearson et al. (1999) found that MC items elicited a significantly lower proportion of multiple and inter-textual strategies than OE items.

Ozuru, Best, Bell, Witherspoon, and McNamara (2007) conducted experiments on the influence of test formats on reading comprehension performance. They varied question formats, i.e. MC vs. OE items and passage availability, i.e. allowing the test taker to access the text while answering comprehension questions (with-text condition) or taking the text away (without-text condition). While the authors found high and significant correlations between the test takes' performance when answering MC items and OE items in the without-text condition, there were only very low and non-significant correlations in the with-text condition. Ozuru et al. (2007) deduced that "the processes underlying OE and MC format item answering in the with-text condition are likely to share less similarity [than those in the without-text condition]" (p. 426).

The studies outlined generally found that MC items in reading assessment seem to be easier than OE items. Shohamy (1984), Pearson et al. (1999), and Ozuru et al. (2007) further suggested differences in underlying reading processes, but only Pearson et al. (1999) further investigated the nature of these differences. One way to approach differences in answering MC and OE items and underlying reading processes is to look for differential correlations to reading precursor skills like general cognitive abilities, vocabulary knowledge, orthography knowledge and reading fluency.

General cognitive abilities are known to be positively related to reading test performance (e.g. Artelt, Schiefele, Schneider, & Stanat, 2002). Several publications have described that individual differences in cognitive abilities affect the acquisition and development of lower level reading skills like decoding, i.e. the ability to recognize and decode words (e.g., Hoskyn & Swanson, 2000; Morris et al., 1998; Snow, Burns, & Griffin, 1998; Stuebing et al., 2002).

A review on the role of *vocabulary knowledge* in reading (Sénéchal, Ouellette, & Rodney, 2006) reported moderate correlations between vocabulary and both decoding and reading comprehension. Ouellette (2006) distinguished vocabulary breadth and depth on the one hand and decoding, visual word recognition and reading comprehension on the other hand. She found that vocabulary breadth predicted decoding performance and visual word recognition, while depth of vocabulary knowledge directly predicted reading comprehension.

Fluent reading rests on word recognition and is hence related to *orthography knowledge*. The effect of orthography knowledge on reading was recently investigated with respect to lower level reading skills such as word identification (e.g. Conrad, 2008), word decoding and reading fluency (e.g. Georgiou, Parrila, & Papadopoulos, 2008). Nevertheless, if difficulty is encountered recognizing individual words any higher order skills such as comprehension can operate (Laberge & Samuels, 1974; Snowling, 2000).

Reading fluency is the ability to read text accurately and quickly. Perfetti (1985, 1992) has seen fast operating word identification processes as the foundation for text comprehension. As reading skill develops automatic word recognition subsequently enables the devotion of mental resources to the meaning of a text and thus allows readers to clearly use reading as a tool for the acquisition of new information and knowledge (Perfetti, 1998; Samuels & Flor, 1997; Spear-Swerling & Sternberg, 1994).

Dimensionality of composite assessments

Empirical studies on the dimensionality of reading assessments which combined MC items and OE items are relatively rare (see van Krieken, 1993 for a study on multiple-choice items vs. guided summaries). Nonetheless, investigations on the dimensionality of composite assessments can be found in the area of assessments of computer science (e.g. Bennett, Rock, & Wang, 1991; Thyssen, Wainer, & Wang, 1994), mathematics (e.g. Walker & Beretvas, 2001), science and language (e.g. Ercikan et al., 1998), and economics (e.g. Becker & Johnston, 1999). The literature on the question whether tests comprising MC and OE items are multidimensional is equivocal.

Two reviews on the question of construct equivalence of MC and OE items from the beginning 1990s came to different conclusions. Traub and MacRury (1990) recommended that “test developers should not assume that MC tests measure the same cognitive characteristics as OE tests, independent of whether the latter are of the essay type or discrete item variety” (p.156, translation by the authors). Wainer and Thissen (1993) suggested the existence of an OE factor, too. But as they found the MC and OE factor highly correlated and as the MC factor was always the more reliable measure, they concluded “measuring something that is not quite right accurately may yield far better measurement that measuring the right thing poorly” (p.115).

A meta-analysis by Rodriguez (2003) emphasized the role of design characteristics of test items. The author found that when items were constructed in MC and OE format both using the same item stem (stem-equivalence) the mean correlation between the two formats was close to one and significantly higher than when both item types didn’t use the same stem. Specifically, the mean correlations between the two formats (all corrected for attenuation) were .92 in stem-equivalent designs, .87 in content equivalent designs, and .82 in non-content equivalent designs. Rodriguez (2003) therefore concluded that while stem-equivalent items appeared to measure the same construct this didn’t seem to be the case for items that are not content equivalent and added: “One may wonder why we would then combine scores from MC and CR [constructed responses, added by the authors] items” (p. 180). Following Rodriguez (2003), different scores for MC and OE items should be reported when both item types were not developed to be stem-equivalent.

In terms of IRT measurement these scores could be attained either by simply applying two separate unidimensional models to each set of items or by applying some sort of multidimensional IRT model. Applying two unidimensional IRT models to reading proficiency data doesn’t reflect the relations between abilities needed to solve both types of items. This interdependence can be explicitly modeled in multidimensional IRT models (Briggs & Wilson, 2003).

Multidimensional IRT models

The usefulness of multidimensional item response theory (MIRT) models for coping with measurement problems in large scale assessment has been pointed out since the beginning 1990s (Ackerman, 1992; Camilli, 1992; Embretson, 1991; Glas, 1992; Luecht & Miller, 1992; Oshima & Miller, 1992; Reckase & McKinley, 1991). Nevertheless, the application of MIRT models in practical testing is relatively rare (Adams, Wilson, & Wang, 1997; Hartig & Höhler, 2008).

Adams et al. (1997) tried to overcome existing application problems by introducing the multidimensional random coefficients multinominal logit model (MRCMLM). Within the MRCMLM framework two subclasses of models are distinguished, such with *between-item multidimensionality* and such with *within-item multidimensionality*. In between-item multidimensional models, each dimension in the model is measured by a separate disjunctive cluster of items. These models can also be characterized as having an independent-cluster structure (McDonald, 2000), or a simple structure of loadings in factor analytic terms. Opposed to that, models with within-item multidimensionality contain items related to more than one ability dimension. In the simplest two-dimensional case, models with within-item multidimensionality provide a dimensional structure where all items load on a general dimension, and only some items additionally load on a second, nested dimension.

As Hartig and Höhler (2008) pointed out both classes of models have very different substantive implications. In models with between-item multidimensionality the latent dimensions represent the abilities required for specific groups of items or subscales, and it is not specified "whether these abilities are completely different or share some common elements" (Hartig & Höhler, 2008, p. 92). In models with a nested dimension, the general dimension represents abilities necessary for all items and the nested dimension represents abilities required exclusively for those items which also load on the second dimension. In the case of compound assessment with non-content equivalent item designs the MIRT model with within-item multidimensionality could display abilities needed to answer MC items and OE items in the general factor and explicitly model the unique information added by OE items in the nested latent dimension.

An empirical example of the use of a model with within-item multidimensionality is Walker and Beretvas' (2001) study on the dimensionality of the mathematics component of the Washington Assessment of Student learning (WASL). The authors hypothesized that those mathematical items which require test takers to communicate about mathematics function differentially in favor of test takers that are better able to organize and present their ideas on paper. The focused items required students to explain their thinking using words, numbers or pictures; to describe a graph or a table or explain the way someone else solved (not necessarily correct) a mathematical problem. The authors firstly used differential bundle functioning (DBF) analyses for exploring whether the set of OE items functioned differently in a comparison of students highly capable of expressing their thoughts with those extremely nonproficient. Secondly, a confirmatory factor analyses (CFA) was conducted fitting a one- and a two-dimensional model. The two dimensional model was specified the way that all items loaded on the MC factor, which was interpreted as the main factor measuring mathematical content, and the OE items loaded additionally on a OE factor, named mathematical communication. The two factors were correlated. Results of the DBF and CFA supported the hypotheses that the model with within-item multidimensionality fitted the data best.

Based on the literature review set out in the foregoing sections we will now state three research hypotheses. The first hypothesis concerns the comparison of two alternative IRT models, the second hypothesis specifies expectations on correlations between reading precursor skills and reading comprehension, modeled once as a uni-dimensional and once as a two-dimensional variable, and the third hypothesis concerns the effects of the most prominent grouping variables on reading comprehension again modeled as a uni- and as a two-dimensional variable.

Hypotheses

We want to compare a two-dimensional model of reading test performance with a more traditional, unidimensional one. Like Walker and Beretvas (2001) and Hartig and Höhler (2008) we chose a two-dimensional structure with within-item multidimensionality to analyze the OE items. Our first hypothesis therefore is:

1. A two-dimensional IRT model in which all items load on the first latent dimension, and the open-ended items additionally load on a second dimension, is more appropriate to measure reading proficiency than a unidimensional IRT model.

If a superior fit of the two-dimensional model can be shown, we are further interested in investigating the nature of the two latent dimensions underlying reading test performance. As stated in the foregoing section we expect general cognitive abilities, vocabulary knowledge, orthography knowledge, and reading fluency to be more closely related to the general latent dimension of the two-dimensional model than to the nested latent dimension. Our second research hypothesis therefore is:

2. The correlations between general cognitive abilities, vocabulary knowledge, orthography knowledge, reading fluency and the general latent dimension of the two-dimensional model are higher than the correlations between these variables and the nested latent dimension.

If these differences in correlations are found, we have some support for interpreting the general latent dimension of the two-dimensional model as abilities necessary to master basic reading processes and the nested latent dimension as abilities necessary to master higher reading processes. We are then interested in investigating the diagnostic benefits of a two-dimensional model compared to a unidimensional model. If we are able to show distinct relationships between ability estimates on the latent dimension(s) and external variables, ability profiles could be generated for different groups of test takers. These profiles could then be used to diagnose for members of these groups abilities necessary to master the basic reading processes separated from abilities necessary to master the higher reading processes. The most interesting variables for exploring the diagnostic benefit of a two-dimensional model of reading comprehension are, as argued above, gender, language spoken at home and school track. We therefore state our third research hypothesis:

3. If reading proficiency is measured using a two-dimensional IRT model, differences between girls and boys, students speaking different languages at home, and students from different school tracks will vary between the two dimensions and will also differ from results using a unidimensional model.

Participants

The dataset used in this study is taken from DESI, a large-scale study on 9th grade students' language competencies and language instruction in Germany (Beck & Klieme, 2007; DESI-Konsortium, 2008). In DESI, students' German skills and English as foreign language skills were assessed at the beginning and at the end of school year 2003/2004. A wide range of language proficiency tests were used in the study, for instance reading, writing, grammar, and orthography tests for German and English. The sample for DESI consisted of whole classrooms and was designed to be representative for German ninth-graders. The analyses presented here are based on the assessment of German reading proficiency of 8523 students from 427 classrooms at the end of the school year. Students joint either the Hauptschule (*n* = 1376), the Gesamtschule (*n* = 507), the Realschule (*n* = 2988) or the Gymnasium (*n*=3652). Slightly more girls were participating (males: *n* = 4025; females: *n* = 4498). To identify students which spoke a language other than German at home, it was asked "Which language do you mainly use at home when talking to your parents?" For the sake of brevity, we refer to the students who mainly use a language other than German at home as the "non-German"-group (*n* = 494) in contrast to the "German"-group (*n* = 6567). For fair group comparisons we needed a measure of the socio-economic status (SES). Students' SES was built according to the international socio-economic index (ISEI; Ganzeboom, de Graaf, Treiman, & de Leeuw, 1992). We used the highest ISEI of either a student's father or mother (HISEI; *n* = 5873) as an indicator for that student's SES.

Instrumentation

The reading proficiency test (Willenberg, 2007) consisted of eight texts, four fictional texts (e.g. short-stories) and four non-fictional texts (e.g. newspaper articles), that were accompanied by a total of 38 items. Of the 38 test items, 26 were dichotomously scored MC items. These items typically referred to information which was explicitly stated in the texts. One MC item on the text with the title "the crossopterygian" was for example "Where were most of the living crossopterygians caught?" Students had to choose between the right answer and three distractors, namely a) "in devon", b) "at the Comors", c) "in the sea" and d) "close to Africa". All answers were explicitly mentioned in the text, but only at the Comors more than one living crossopterygian was found. Six items were OE, and asked questions related to the main ideas of the text, the motivation of protagonists' actions, and statements in the text. Some OE items asked students to interpret critical text passages, whose meaning could not be deduced from a literal understanding of the text. A typical OE item asks to describe a group of people mentioned in the text in one's own words. Test takers were allowed between four and ten lines to give rather detailed responses to the OE questions. Responses to OE items were scored dichotomously by trained coders on the basis of standardized coding instructions. Six items required the students to underline relevant passages of texts, to write down numbers related to text parts, to draw a picture, or to give a title to the text (students were allowed three to four words in these cases). However, these six items were excluded from our analysis, since their response formats were too heterogeneous. As the tests were administered in a matrix design, each student had to respond to a subset of the items. On average, each student answered a total of 13 MC and 3 OE items (a total of 16 items). Students had about 25 minutes to read the texts and answer the items. The students had access to the texts while answering the comprehension questions.

General cognitive abilities were measured through the 26 items of the second non-verbal scale of the Cognitive Abilities Test ("Kognitiver Fähigkeitstest", KFT; Heller & Perleth, 2000), which uses figural analogies. In the subtest KFT N2 students have to select one out of five answer alternatives (one correct, four distractors) to complete a pair of figures in analogy to a given example. The KFT showed an EAP/PV reliability of .87. The EAP/PV reliability is an estimate for test reliability that is provided by the ConQuest software (Wu, Adams, & Wilson, 2007) and that is obtained by dividing the variance of the individual expected a posteriori ability estimates by the estimated total variance of the latent ability.

The vocabulary test applied in DESI (Willenberg, 2007) consisted of 44 items, which either required students to label parts of pictures, fill in gaps in texts or find synonyms. In terms of Ouellettes' (2006) distinction the DESI test mainly concentrated on measuring breadth of vocabulary, while assessing vocabulary depth to a minor extent. The DESI vocabulary test had a reliability (EAP/PV) of .73.

Orthography was tested with a dictation that consisted of 68 words (Thomé & Gomolka, 2007). A phase model of literary language acquisition allowed for distinguishing types of errors (e.g. using basic graphemes instead of orthographic graphemes and comma placement). Scoring was based on the occurrence of each error type. The orthography test showed a reliability (EAP/PV) of .74.

For the reading fluency test students had to read as much as possible of a text of 1130 words test in three minutes. In each passage a bracket with three alternative words was included (in sum 12 brackets) and students had to decide which of the presented words fitted best in the given context. Reading fluency then was measured by the number of right choices, where no choice was taken as wrong. The reading fluency test had a reliability (EAP/PV) of .67.

For the KFT, the vocabulary test, the orthography test and the reading fluency test, weighted likelihood estimates (WLEs; Warm, 1989) obtained from unidimensional scaling with ConQuest were used as measures of students' performance.

Results

To address hypothesis 1, we compared the results of the unidimensional model (model 1) and the two-dimensional model (model 2) by testing difference in the log likelihood via χ^2 test and additionally inspecting information criteria (AIC, BIC, sample-size adjusted BIC). To test hypothesis 2 we compared the correlations between reading proficiency, modeled once as a unidimensional variable (model 3) and once as two-dimensional variable (model 4) and general cognitive abilities, vocabulary knowledge, orthography knowledge, and reading fluency. To approach hypothesis 3 we included gender, language spoken at home, and school track as predictors in three latent regression models. The criterion in these regression analyses was reading proficiency, again modeled once as a unidimensional variable (model 5.1-3) and once as two-dimensional variable (model 6.1-3). Herewith we aimed comparing proficiency profiles and evaluating the diagnostic benefit of the multidimensional approach.

All models were analyzed with MPlus 5 (Muthén & Muthén, 2007), using maximum likelihood estimation with robust standard errors. The cluster structure of the sample (students nested within classes) was taken into account by using the pseudo-maximum-likelihood estimator for complex samples implemented in Mplus (Asparouhov & Muthén, 2005). Sampling weights were used to adjust for unequal sampling probabilities.

Hypothesis 1: Comparison of the uni- and two-dimensional IRT model

To address hypothesis 1 we first applied a unidimensional logistic item response model (model 1) with one latent dimension common to all 32 items (26 MC and 6 OE items) to the reading test data. All item-loadings on the latent dimension were fixed to one, which makes the model a one-parameter logistic (1PL) or Rasch model (e.g. Kubinger, 2005). After that, we applied a two-dimensional model (model 2) in which all items are loading on the first latent dimension, and the OE items additionally load on a second dimension. In this kind of within-item multidimensional model (Adams et al., 1997), performance in complex tasks is split into more basic abilities, thus providing a more detailed picture of the competence assessed (Hartig & Höhler, 2008). All loadings of the items on both dimensions were fixed to one as in the unidimensional model. Both the correlation between the two latent dimensions and the variances of the two latent dimensions were freely estimated.

Figure 1 illustrates the models 1 and 2.

Since Mplus doesn't provide fit indices for individual items, the fit for single items was estimated using the ConQuest item response modeling software (Wu et al., 2007). For the unidimensional model, the fit (weighted mean squares) provided by ConQuest ranged from 0.85 to 1.07, for the two-dimensional model from 0.79 to 1.19, thus indicating a good fit of both models on item level. Results of the model comparison are shown in table 1.

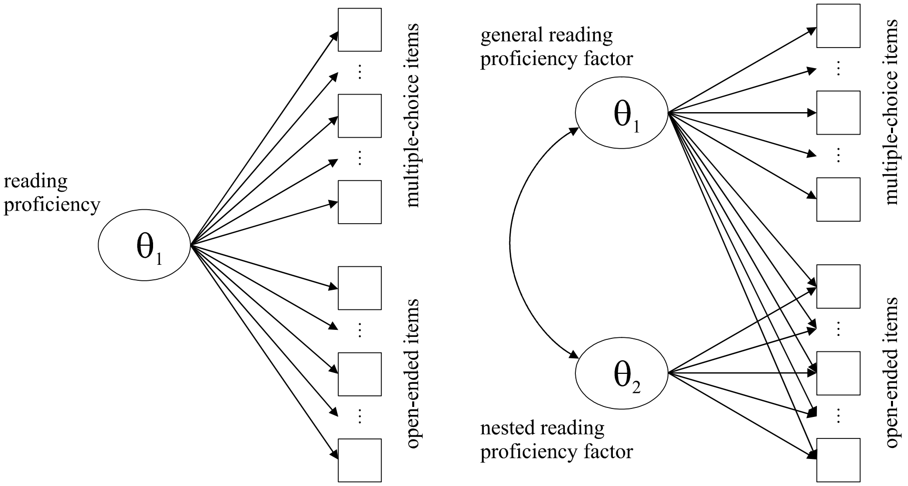


Figure 1:
Graphical illustration of the unidimensional (model 1, left) and the two-dimensional (model 2, right) IRT model applied to the reading proficiency data

The global fit of the unidimensional model is worse than that of the two-dimensional within-item model (χ^2_{Diff} = 7,916; df_{Diff} = 2; $p \leq .05$)[3], which reinforces the assumption that a two-dimensional structure of the data is justifiable. This is furthermore supported by the relatively high variance ($\hat{\sigma}^2 = 0.852$) of the nested OE items latent dimension when compared to the common latent dimension in the two-dimensional model ($\hat{\sigma}^2 = 0.534$) and in the unidimensional model ($\hat{\sigma}^2 = 0.636$). The latent correlation of the two latent dimensions in the two-dimensional model is positive but only moderate (r = .443).

[3] The χ^2_{Diff}-value for the maximum likelihood estimator with robust standard errors was calculated according to the procedure suggested by Satorra and Bentler (1999; see http://www.statmodel.com/chidiff.shtml) and takes into account a scaling correction factor provided by Mplus. The scaling correction factors were 1.401 for the unidimensional model (model 3) and 1.371 for the two-dimensional model (model 4); the corresponding loglikelihoods are reported in Table 1.

Table 1:
Number of free parameters and global fit indices for the unidimensional model and the two-dimensional within-item model (N =8523)

Model	**Free parameters**	**LL**	**AIC**	**BIC**	**Sample-size adjusted BIC**
Unidimensional (model 1)	33	-78888	157842	158075	157970
Two-dimensional (model 2)	35	-78614	157298	157545	157433

Notes: model 1.1 = unidimensional IRT model; model 1.2 = two-dimensional IRT model; LL = Log-likelihood; AIC = Akaike's information criterion; BIC = sample-size-adjusted Bayesian information criterion.

Hypothesis 2: Correlations of reading precursor skills with latent dimension(s)

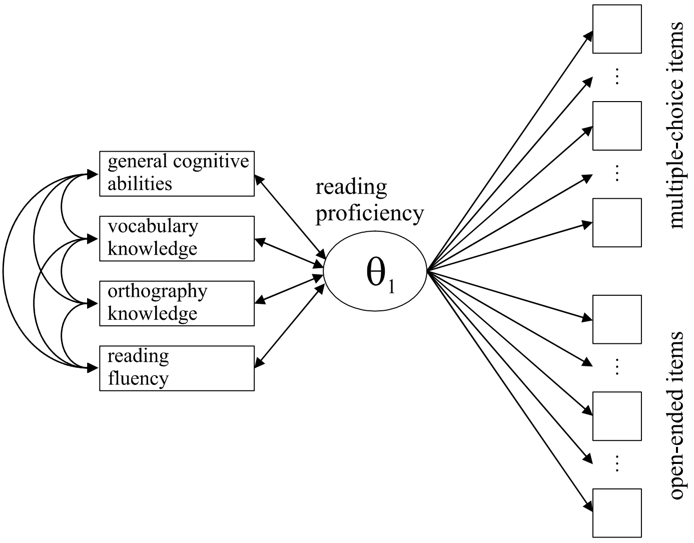
To test our second hypothesis, we estimated the correlations between reading proficiency modeled as a two-dimensional variable and general cognitive abilities, vocabulary knowledge, orthography knowledge and reading fluency. To allow for comparisons, we first applied a unidimensional model to the reading test data, now additionally estimating the correlations between reading proficiency and each of the other tests (model 3). Secondly we applied the two-dimensional within model and estimated correlations between the general latent dimension and the other tests on the one hand and the nested latent dimension and the other tests on the other hand (model 4). Figure 2 illustrates the resulting models 3 and 4.

To test the correlation differences postulated in Hypothesis 2, these differences were defined as additional parameters to be estimated in Mplus for each of the relevant variables (cognitive abilities, vocabulary knowledge, orthography knowledge, reading fluency). For each variable X, the difference $\Delta\rho_X$ between the correlation with the general latent dimension θ^G and with the nested latent dimension θ^N was defined based on the variances and covariances that are estimated by default:

$$\Delta\rho_X = \frac{\text{cov}\left(\theta^G, X\right)}{\sqrt{\text{var}\left(\theta^G\right)}\cdot\sqrt{\text{var}\left(X\right)}} - \frac{\text{cov}\left(\theta^N, X\right)}{\sqrt{\text{var}\left(\theta^N\right)}\cdot\sqrt{\text{var}\left(X\right)}}$$

Thereby, for each variable relevant for hypothesis 2, the correlation difference $\Delta\rho_X$ was tested for statistical significance. Table 2 shows the correlations estimated in models 3 and 4 and the correlation differences.

All correlations and all differences in correlations were found to be significant. The pattern of correlations between the two latent dimensions of the two-dimensional model and the other covariates was as expected in three of four cases. General cognitive abilities, vocabulary knowledge and orthography knowledge correlated significantly and substantially higher with the general latent dimension than with the nested latent dimension. However, the difference in the correlations between reading fluency and the general and the nested latent dimension was comparatively low.



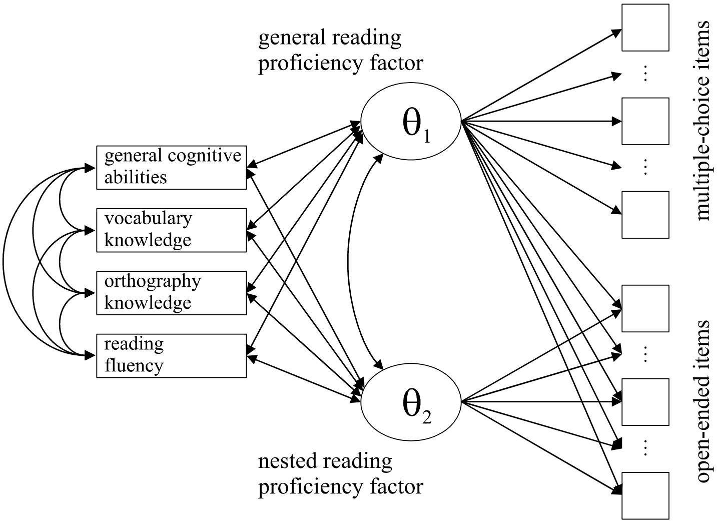


Figure 2:
Graphical illustration of the correlations of general cognitive abilities, vocabulary knowledge,orthography knowledge and reading fluency with reading proficiency (unidimensional IRT model (model 3) above, two-dimensional IRT model (model 4) below)

Table 2:
Correlations between reading proficiency (uni-dimensional and two-dimensional IRT model) and general cognitive abilities, vocabulary knowledge, orthography knowledge, and reading fluency (standard errors in brackets).

	Model 3	Model 4		
	reading proficiency	**reading proficiency general latent dimension**	**reading proficiency nested latent dimension**	**differences between correlations in model 4**
general cognitive abilities	0.583 (0.017)	0.601 (0.017)	0.256 (0.031)	0.344 (0.034)
vocabulary knowledge	0.706 (0.014)	0.710 (0.016)	0.411 (0.029)	0.299 (0.032)
orthography knowledge	0.526 (0.016)	0.526 (0.017)	0.315 (0.027)	0.210 (0.031)
reading fluency	0.410 (0.019)	0.398 (0.020)	0.308 (0.026)	0.089 (0.031)

Notes: model 3 = unidimensional IRT model; model 4 = two-dimensional IRT model.

Hypothesis 3: Latent regression analysis and proficiency profiles

To approach our third hypothesis six models with latent regressions were estimated: the regressions of unidimensional reading proficiency on gender (model 5.1), on language spoken at home (model 5.2), and on school track (model 5.3) and the effects of the general and nested latent dimensions in a two-dimensional model on gender (model 6.1), on language spoken at home (model 6.2), and on school track (model 6.3).

School track is split up in three dummy variables with Hauptschule as reference category and gender and language spoken at home are coded 0/1 for male/female and German/non-German. Regression coefficients (β) were standardized on basis of the standard deviation of the latent dimensions. Regression coefficients therefore display the latent mean difference between males and females, respectively German speaking students and non-German speaking students, and Hauptschule and each of the other three school tracks. As language spoken at home is highly confounded with the school track and the parents' socioeconomic status it seems reasonable to add regression analysis in which these covariates are controlled for (amending models 5.2 and 6.2). Table 3 displays the standardized coefficients of three latent regressions: the regressions of reading proficiency on gender, on language spoken at home and on school track.

In model 5.1 the effect of gender amounts to a little under two tenth of a standard deviation of the variation in the reading test performance, making gender only a moderate but still significant predictor of performance in the reading test. Conversely, the negative effect of speaking a language other than German at home (model 5.2) is half of a stan-

Table 3:
Regression coefficients (β), standardized on basis of SD of criterion, from the latent regressions of reading test performance (unidimensional and two-dimensional) on gender (N=8523), language spoken at home (N=8523), and school track (N = 8523) (standard errors in brackets).

Predictor	**Unidimensional model**	**reading proficiency**	**Two-dimensional model**	**reading proficiency general latent dimension**	**reading proficiency nested latent dimension**
Gender: female	Model 5.1	0.190 (0.041)	Model 6.1	0.112 (0.043)	0.542 (0.058)
Language spoken at home: not German	Model 5.2	-0.514 (0.075)	Model 6.2	-0.463 (0.081)	-0.636 (0.130)
School track:	Model 5.3		Model 6.3		
Gesamtschule		0.393 (0.133)		0.422 (0.135)	0.105 (0.144)
Realschule		0.764 (0.057)		0.770 (0.058)	0.460 (0.085)
Gymnasium		1.640 (0.046)		1.659 (0.049)	0.845 (0.084)

dard deviation of the reading test, which is quite substantial. When applying the unidimensional model, the greatest effect of the school track variable can be found when comparing students from the Gymnasium with those from the Hauptschule (model 5.3). However, a noticeable effect can already be found when comparing the reading test performance of students from the Gesamtschule and the Realschule with those from the Hauptschule.

The comparison with the second set of regression analysis reveals that the effects of gender, language spoken at home, and school track vary substantially between the unidimensional model and the two-dimensional model, and also between both dimensions within the two-dimensional model. The most striking difference between the two modeling approaches can be found when looking at the effect of gender (model 6.1). While there is a significant, moderately positive, gender difference in favour of girls in reading proficiency modeled as a single dimension, there is no significant effect of gender on the general reading proficiency dimension in the two-dimensional model. However, gender shows a strong relationship with the nested dimension of the two-dimensional model. The effects of language spoken at home on the general and nested dimensions in the two-dimensional model (model 6.2) differ about twenty per cent of a standard deviation of reading proficiency. When school track and SES are controlled for, students who speak another language than German at home perform not significantly worse than other students when reading proficiency is modeled unidimensional (β= -0.094). In the two-dimensional model the control of these covariates reveals that the effect of language at home is non-significant for the general reading proficiency dimension (β =-0.040), too, but significant and quite substantial for the nested dimension (β =

-0.352). While the effect of the school track on the general dimension of reading proficiency is almost the same as in the unidimensional model, this effect is noticeably lower for the nested dimension of reading proficiency (model 6.3). Students from Gymnasium perform over one and a half standard deviations better on the general reading proficiency dimension than students from the Hauptschule, this difference is reduced by half for the nested reading proficiency dimension. The effect of visiting a Gesamtschule in comparison to a Hauptschule on the nested dimension is not significant and only a quarter of the effect on the general dimension.

Discussion

Hypothesis 1: Comparison of the uni- and two-dimensional IRT model

The results of this study suggested that reading proficiency measured simultaneously with MC and OE items can be described more adequately with a two-dimensional IRT model than with a unidimensional model. In the two-dimensional model, proficiency aspects which all items have in common build a general latent dimension and proficiency aspects specific to OE items build an additional, nested latent dimension. We were further able to show that a substantial amount of variance in reading test performance is due to differences in ability that are only measured by OE items. With this we are able to replicate findings from L2 assessments (Shohamy, 1984; Kobayashi, 2002; Wolf, 1993) with data from L1 reading assessment. This may indicate that effects of response formats in reading may be generalized across reading assessments in different languages.

It is important to realize that possible generalizations of our findings are limited to reading assessments with non-stem equivalent multiple-choice and open-ended items, since the DESI MC items were substantially differently constructed than the OE items (see *instrumentation* for an example). As Rodriguez' (2003) meta-analysis showed for a variety of assessed skills, stem-equivalent items are more likely to measure the same construct.

A shortcoming of the data and analyses presented here is that item format and reading processes are confounded in the way the DESI reading proficiency test was constructed – open items were generally intended to assess higher level processes compared to MC items. Theoretically, MC items could also be constructed to assess abilities necessary to master higher reading processes, and OE-items don't necessarily assess abilities necessary to master higher reading processes. The confounding of format and cognitive processes could be avoided if reading items were systematically constructed to keep item format and reading process independent – e.g. by explicitly instructing item writers to assess abilities necessary to master higher reading processes with open ended as well as with closed response formats. However, for item writers it is often easier to construct items assessing abilities necessary to master higher reading processes with open response formats and to construct items aiming at abilities necessary to master basic reading processes with closed response formats. Thus, response format and assessed skills and cognitive processes are very likely to be confounded in applied assessments.

Hypothesis 2: Correlations of reading precursor skills with latent dimension(s)

As expected, we could show that general cognitive abilities, vocabulary knowledge and orthography knowledge correlated significantly and substantially higher with the general dimension of the two-dimensional model than with the nested dimension. Based on research on the relation of these reading comprehension precursor skills and basic reading abilities like decoding and word recognition, we suggested that this general dimension can be interpreted as abilities necessary to master basic reading processes that are needed for solving both MC and OE items. We further suggested that the nested dimension, which is measured only with OE items, tests abilities necessary to master higher reading processes.

Answers to multiple-choice and open-ended items in reading comprehension assessments can be influenced by other, more formal aspects of item formats. Therefore, and given the confounding of response format and reading processes mentioned above, the suggested interpretations of the latent dimensions have to stand against other possible interpretations, the most appealing among them are probably 1) active language skills necessary to formulate an answer, 2) guessing, and 3) test taking strategies.

1. For the OE items, several other skills than reading are required to formulate an appealing answer (vocabulary, writing skills, etc.) that cannot be equated with higher reading comprehension skills. Referring to the coding-guidelines in the DESI-study, this interpretation however seems less sensible. Answers only of three to four words long where scored correct if only they met the expectation content wise; orthography didn't matter at all as long as the answers were understandable.
2. For the MC items, the students have the possibility to guess which is not possible for the open-ended items. If students differ in the degree to which they rely on guessing, this may affect unidimensionality of the reading test. While this interpretation cannot be ruled out completely with the given data, it doesn't seem too plausible given the distinct correlations of the two dimensions with language specific variables like vocabulary and orthography knowledge. Additionally, the relatively high variance of the nested latent dimension specific to the OE items indicates that the lack of unidimensionality is due to performance variation specific to the OE items rather than to the MC items.
3. Test-taking strategies related to testing time limits might also explain the lack of unidimensionality: Since test-takers consider OE questions to be more challenging and more time consuming than MC items, they might focus on answering MC items first. If students differ in the extent to which they apply this strategy, this might introduce multidimensionality, too. These strategies become more relevant if the time limit is narrow. However, students had enough time to read the texts several times and to answer all questions which they felt capable of. Therefore, the influence of time-related test-taking strategies can be expected to be rather low.

Hypothesis 3: Latent regression analysis and proficiency profiles

Finally, we were able to demonstrate how the two-dimensional IRT model could be used to build group profiles of reading proficiency in order to learn more about the specific weaknesses and strengths of these groups of students. As we could see from the comparison of the standardized regression coefficients of gender in both IRT models, the differences between girls and boys were not due to abilities necessary to master lower reading processes but were caused by abilities necessary to master higher reading processes, measured with OE items. As with the effect of gender, the effect of the language spoken at home was not the same within both IRT models and on both dimensions in the two-dimensional IRT model. As we learned from the two-dimensional model, speaking a language other than German at home is a disadvantage for abilities necessary to master basic reading processes, and it turned out be even a greater disadvantage for abilities necessary to master higher reading processes. When controlling for SES and school track, the effect of speaking another language than German at home vanished for abilities necessary to master basic reading processes, while there was still a remarkably negative effect on abilities necessary to master higher reading processes. We further found the advantages of students from Gesamtschule, Realschule and Gymnasium over students from the Hauptschule to be smaller in abilities necessary to master higher reading processes than in abilities necessary to master basic reading processes.

Limitations

The within-item multidimensional IRT model applied in the present study is compensatory. It assumes that the abilities necessary to master basic reading processes measured with MC items are also needed to answer OE items successfully. In other words, with regard to test takers working on the OE items, a lack of abilities necessary to master basic reading processes can be compensated by strong abilities necessary to master higher reading processes and vice versa. From a theoretical perspective, this has to be considered carefully: lacking abilities necessary to master basic reading processes can be compensated by in-depth knowledge of text content (Schneider, Körkel, & Weinert, 1989; Voss & Silfies, 1996); Nonetheless, this applies only to the level of the situation model and not on the level of the propositional text base (Moravcsik & Kintsch, 1993). With regard to OE questions focusing on the building of a situation model, it is possible to compensate for abilities necessary to master basic reading processes, by constructing meaning from extended prior knowledge (Bisanz, Das, Varnhagen, & Henderson, 1992). In contrast to this, compensating the lacking ability to incorporate prior knowledge into a coherent situation model through excellent abilities necessary to master basic reading processes is improbable, as is compensation in general when extreme ability levels are considered. A total lack of either ability would be expected to lead to a failure in the OE items.

However, for positively correlated dimensions, the predicted response probabilities are quite similar for a compensatory and a non-compensatory model (Hartig & Höhler, 2009). Since ability dimensions were positively correlated in our study, the application of a non-compensatory model would probably lead to results comparable to those of the compensatory model.

It is certainly only a loose link that we established between cognitive processes and cognitive representations and text challenges posed by item formats in reading proficiency assessment. Only for three of four reading precursor skills the expected higher correlation with the general dimension was found. For reading fluency the difference in correlations with the general and the nested dimension was quite small. Therefore the correlations with reading fluency didn't truly help for specifying the abilities needed for the general and the nested dimension. Moreover, it would have been preferable to additionally show reverse correlation patterns with another set of variables which could be meaningfully related to the nested dimension thereby strengthening its interpretation as abilities necessary to master higher reading processes.

In general, as Davey (1987, 1988) pointed out, the numerical score obtained in a standardized reading proficiency assessment provides no in-depth insight into the challenges that test takers must confront and master in order to answer standardized comprehension questions successfully. What is required to effectively close this gap and to apply results from experimental research to the relation of item format and cognitive processes in reading comprehension is a systematic variation of the content of questions and, hence, of the associated cognitive processes and cognitive text representation required on one hand and item format on the other hand.

Consequences for teaching and testing

Based on our results the common findings of higher reading proficiency of girls is due to reading proficiency aspects often measured with OE items in standardized reading tests. Of course, if these abilities form part of the test construct of standardized reading proficiency tests, as is suggested in the reading proficiency definitions of prominent studies such as PISA and PIRLS, the different performances of boys and girls in OE items can not be interpreted as a method factor but have to be interpreted substantively. Our results suggested that it is a particular weakness of boys compared to girls to integrate text and prior knowledge to build an adequate situation model. To teach reading in a way that accounts for this need therefore could mean to encourage boys to state their understandings of texts explicitly in the classroom and to motivate them by choosing themes of high interest and prior knowledge. Taking into account the much smaller effect of language spoken at home on the basic reading ability dimension, the effect on abilities needed to master higher reading processes cannot be explained satisfactorily on grounds of language difficulties only. The students referred to are disadvantaged when asked to interpret text passages by incorporating personal experience and previous knowledge; this could be down to a pronounced difference in their knowledge and experience base (Bühler-Otten, Neumann, & Reuter, 2000) or it might be caused by cultural references in the texts that are different to those the students are familiar with. A tailored instruction in class that bridges these gaps could thus help improving the reading proficiency of nonnative speakers. These classes should address abilities necessary to master lower reading processes too, but should not merely train e.g. how to elicit explicitly stated propositions from a text. Surprisingly, for abilities necessary to master higher reading processes the effects of school track were smaller than for abilities necessary to master lower reading processes. This could mean that school track first and foremost is related to these basic abilities, while higher abilities do not depend so much on school track. Reading comprehension instruction is of course virtually completed at the age of 15, especially instruction aiming at improving abilities necessary to master lower reading processes. It might even be that students are in parts separated due to reading abilities after completing the primary school. In this case causality would be the other way round. Nevertheless, teachers at Hauptschule, Gesamtschule and Realschule should concentrate on facilitating lower reading processes in order to catch up to the Gymnasium.

Our study underlined the need to the improve the diagnostic value of reading comprehension tests, in which MC items and OE items are not stem-equivalent, by using a two-dimensional IRT model with within-item-multidimensionality and reporting two scores. When considering the consequences for scoring, we have to take one restriction into account: the DESI study, like most studies, contains too few OE items to be used as basis for a reliable measuring of the nested dimension. In contrast to Wainer and Thissen (1993), we nevertheless wouldn’t conclude that therefore a unidimensional scaling is more appropriate, but that whenever a multidimensional scaling of reading proficiency tests is intended, more OE items have to be included.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

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

Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.

Alderson, C. J., Figueras, N., Kuijper, H., Tardieu, C., Nold, G., & Takala, S. (n.d.): The Dutch CEFR Grid Reading/Listening. Retrieved April 9, 2009, from http://www.lancs.ac.uk/fss/projects/grid.

Artelt, C., Schiefele, U., Schneider, W., & Stanat, P. (2002). Leseleistungen deutscher Schülerinnen und Schüler im internationalen Vergleich: Ergebnisse und Erklärungsansätze [Reading performance of German students in international comparison: results and explanation attempts]. *Zeitschrift für Erziehungswissenschaft, 5,* 6-27.

Artelt, C., Stanat, P., Schneider, W., & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse [Reading competence: test concept and results]. In Deutsches PISA-Konsortium (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* [PISA 2000. Basic student competencies in international comparison.] (pp. 69-140). Opladen, Germany: Leske + Budrich.

Asparouhov, T., & Muthén, B. (2005). *Multivariate Statistical Modeling with Survey Data.* Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference.

Beck, B., & Klieme, E. (Eds.) (2007). *Sprachliche Kompetenzen. Konzepte und Messun*g [Language competence. Concepts and measurement]. Weinheim, Germany: Beltz.

Becker, W. E., & Johnston, C. (1999).The Relationship between Multiple Choice and Essay Response Questions in Assessing Economics Understanding. *Economic Record, 75,* 348-357.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28,* 77-92.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats – It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*, 385-395.

Bisanz, G. L., Das, J. P., Varnhagen, C. K., & Henderson, H. R. (1992). Structural components of reading times and recall for sentences in narratives: Exploring changes with age and reading ability. *Journal of Educational Psychology, 84*, 103-114.

Briggs, D. C., & Wilson, M. (2003). An Introduction to Multidimensional Measurement using Rasch Models. *Journal of Applied Measurement*, *4,* 87-100.

Brown, J. I., Vick-Fishco, V., & Hannah, G. S. (1993). *Nelson-Denny Reading Test*. Rolling Meadows, IL: Riverside Publishing.

Bühler-Otten, S., Neumann, U., & Reuter, L. (2000). Interkulturelle Bildung in den Lehrplänen [Intercultural education in school curricula], In I. Gogolin & B. Nauck (Eds.), *Migration, gesellschaftliche Differenzierung und Bildung* [Migration, differentiation of society and education] (pp. 279-320). Opladen, Germany: Leske + Budrich.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16,* 129-147.

Conrad, N. J. (2008). From reading to spelling and spelling to reading: transfer goes both ways. *Journal of Educational Psychology, 100,* 869-878.

Cordon, L. A., & Day, J. D. (1996). Strategy use on standardized reading comprehension tests. *Journal of Educational Psychology, 88,* 288-295.

Daneman, M., & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology, 11,* 175-193.

Davey, B. (1987). Postpassage questions: Task and reader effects on comprehension and metacomprehension processes. *Journal of Reading Behavior, 19,* 261-278.

Davey, B. (1988). The nature of response errors for good and poor readers when permitted to reinspect text during question-answering. *American Educational Research Journal, 25,* 399-414.

DESI-Konsortium (Eds.) (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* [Teaching and the development of expertise in German and English. Results of the DESI study]. Weinheim, Germany: Beltz.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495-515.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35,* 137-154.

Ganzeboom, H. B. G., de Graaf, P. M., Treiman, D. J., & de Leeuw, J. (1992): A standard international socio-economic index of occupational status. *Social Science Research, 21,* 1-56.

Georgiou, G. K., Parrila, R., & Papadopoulos, T. C. (2008). Predictors of word decoding and reading fluency across languages varying in orthographic consistency. *Journal of Educational Psychology, 100,* 566-580.

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective Measurement: Theory into practice* (pp. 236-258). Norwood NJ: Ablex.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie / Journal of Psychology, 216,* 88-100.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35,* 57-63.

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen*. [Test of cognitive abilities for grades 4-12] Revision. Göttingen, Germany: Hogrefe.

Hoskyn, M., & Swanson, H. L. (2000). Cognitive processing of low achievers and children with reading disabilities: A selective meta-analytic review of the published literature. *School Psychology Review, 29,* 102-119.

Johns, J. L. (1978). Do Comprehension Items Really Test Reading? Sometimes! *Journal of Reading, 21,* 615-619.

Katz, S., Lautenshalger, G. J., Blackburn, A. B., & Harris, F. H. (1990). Answering reading comprehension items without passages on SAT. *Psychological Science, 1,* 122-127.

Kim, S., Walker, M. E., & McHale, F. (2009). Equating of mixed-format tests in large scale assessments. *ETS Research Spotlight, 2,* 15-20.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, UK: Cambridge University Press.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review 85,* 363-394.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Language Testing, 19,* 193-220.

van Krieken, R. (1993). *Construct validation of question formats for Dutch central examination in foreign language reading comprehension.* (Reports – research/ technical 143). Arnhem, the Netherlands: Cito.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model – Some critical suggestions on traditional approaches. *International Journal of Testing, 5,* 377-394.

Laberge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6,* 293-323.

Luecht, R. M., & Miller, R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16,* 279-293.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24,* 99-114.

Moravcsik, J. E., & Kintsch, W. (1993). Writing quality, reading skills, and knowledge as factors in text comprehension. *Canadian Journal of Experimental Psychology, 47,* 360-374.

Morris, R. D., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Steubing, K. K., Fletcher, J. M., et al. (1998). Subtypes of reading disability: Variability around a phonological core. *Journal of Educational Psychology, 90,* 347-373.

Mullis, I. V. S., Kennedy, A. M., Martin, M. O., & Sainsbury, M. (2006). *PIRLS 2006. Assessment framework and specifications* (2nd ed.). TIMSS & PIRLS International Study Center. Lynch School of Education, Boston College.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

OECD. (2001). *Knowledge and skills for life: First results from the OECD Program for International Student assessment (PISA) 2000*. Paris: OECD.

OECD. (2003). *The PISA 2003 assessment framework – mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD.

OECD. (2004). *Learning for Tomorrows world: First results from PISA 2003*. Paris: OECD.

OECD. (2006). *Where immigrant students succeed – a comparative review of performance and engagement in PISA 2003*. Paris: OECD.

OECD. (2007) *Executive Summary PISA 2006: Science Competencies for Tomorrow's World*. Paris: OECD.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16,* 237-248.

Ouellette, G. P. (2006). What's meaning got to do with it: the role of vocabulary in wordreading and reading comprehension. *Journal of Educational Psychology, 98,* 554-566.

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction, 25,* 399-438.

Pearson, P. D., Garavaglia, D., Lycke, K., Roberts, E., Danridge, J., & Hamm, D. (1999). *The impact of item format on the depth of students' cognitive engagement*. Washington, DC: Technical Report, American Institute for Research.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Perfetti, C. A. (1992). The representation problem in reading acquisition. In P. Gough, L. Ehri, & R. Trieman (Eds.), *Reading acquisition* (pp.145-174). Mahwah, NJ: Erlbaum.

Perfetti, C. A. (1998). Learning to read. In P. Reitsma & L. Verhoeven (Eds.), *Literacy problems and interventions* (pp. 15-48). Dordrecht: Kluwer.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361-373.

Rodriguez, M. C. (2003) Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations. *Journal of Educational Measurement, 40,* 163-184.

Samuels, S. J., & Flor, R. (1997). The importance of automaticity for developing expertise in reading. *Reading and Writing Quarterly, 13,* 107-122.

Satorra, A., & Bentler, P. M. (1999). *A scaled difference chi-square test statistic for moment structure analysis.* (UCLA Statistics Series No. 260). Retrieved April 9, 2009, from http://preprints.stat.ucla.edu/260/chisquare.pdf

Schneider, W., Körkel, J., & Weinert, F. E. (1989). Domain-specific knowledge and memory performance: A comparison of high- and low-aptitude children. *Journal of Educational Psychology, 81,* 306-312.

Sénéchal, M., Ouellette, G., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. In S. B. Neuman & D. Dickinson (Eds.), *Handbook of early literacy research*: Vol. 2 (pp. 173-182). New York: Guilford Press.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1,* 147-170.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Snowling, M. J. (2000). Language and literacy skills: Who is at risk and why? In D. V. M. Bishop & L. B. Leonard (Eds.), *Speech and language impairments in children: Causes, characteristics, intervention and outcome* (pp. 245-259). New York: Psychology Press.

Spear-Swerling, L., & Sternberg, R. J. (1994). The road not taken: An integrative theoretical model of reading disability. *Journal of Learning Disabilities, 27*, 91-103.

Stanat, P., Rauch, D. P., & Segeritz, M. (2010). Schülerinnen und Schüler mit Migrationshintergrund. [Students with migrational background] In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider & P. Stanat (Hrsg.), *PISA 2009. Bilanz nach einem Jahrzehnt* [PISA 2009. Conclusions after a decade.] (S. 200-230). Waxmann, Germany: Münster.

Stuebing, K. K., Fletcher, J. M., LeDuox, J. M., Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2002). Validity of IQ-discrepancy classifications of reading disabilities: A meta-analysis. *American Educational Research Journal, 39,* 469–518.

Thomé, G., & Gomolka, J. (2007). Rechtschreibung [Orthography], In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung* [Language competencies. Concepts and measurement] (pp.140-146). Weinheim, Germany: Beltz.

Thyssen, D., Wainer, H., & Wang, X.-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31,* 113-123.

Traub, R. E., & MacRury, K. (1990). Multiple-choice vs. free-response in the testing of scholastic achievement. In K. Ingenkamp & R.S. Jager (Eds.), *Tests und Trends. 8. Jahrbuch der Pädagogischen Diagnostik* [Test and trends. 8th annual of educational diagnostic] (pp. 128-159). Weinheim, Germany: Beltz.

Voss, J. F., & Silfies, L. N. (1996). Learning from history text: The integration of knowledge and comprehension skill with text structure. *Cognition and Instruction, 14,* 45-68.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education 6,* 103-118.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement, 38,* 147-163.

Warm T. A. (1989).Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika, 54,* 427-450.

Willenberg, H. (2007). Lesen [Reading]. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung* [Language competencies. Concepts and measurement] (pp. 107-117). Weinheim, Germany: Beltz.

Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *The Modern Language Journal, 77,* 473-489.

Wu, M. L., Adams, R. J., & Wilson, M. R. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software.* Melbourne: ACER Press.

Zwaan, R. A., & Singer, M. (2003). Text comprehension. In A. C. Grasser, M. A. Gernsbacher & S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 83-121) Mahwah, NJ: Lawrence Erlbaum Associates.

Acknowledgement

The DESI study was funded by the German Standing Conference of the Ministers of Education and Cultural Affairs of the Länder (Kultusministerkonferenz, KMK).

The preparation of this paper was partly supported by grant HA 5050/2-2 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).