Frontline Learning Research

An official journal of EARLI

Frontline Learning Research Vol.3 No. 2 (2015) 1-26
ISSN 2295-3159

# Competence Assessment of Students With Special Educational Needs—Identification of Appropriate Testing Accommodations

**Anna Südkamp[a], Steffi Pohl[b], & Sabine Weinert[c]**

[a]TU Dortmund University, Germany

[b]Freie Universität Berlin, Germany

[c]University of Bamberg, Germany

*Article received 23 October 2014 / revised 16 March 2015 / accepted 18 May 2015 / available online 1 June 2015*

## Abstract

*Including students with special educational needs in learning (SEN-L) is a challenge for large-scale assessments. In order to draw inferences with respect to students with SEN-L and to compare their scores to students in general education, one needs to assure that the measurement model is reliable and that the same construct is measured for different samples and test forms. In this article, we focus on testing the appropriateness of competence assessments for students with SEN-L. We specifically asked how the reading competence of students with SEN-L may be assessed reliably and comparably. We thoroughly evaluated different testing accommodations for students with SEN-L. The reading competence of N = 433 students with SEN-L was assessed using a standard reading test, a reduced test version, and an easy test version. Also, N = 5,208 general education students and a group of N = 490 low-performing students were tested. Results show that all three reading test versions are suitable for a reliable and comparable measurement of reading competence in students without SEN-L. For students with SEN-L, the accommodated test versions considerably reduced the amount of missing values and resulted in better psychometric properties than the standard test. They did not, however, show satisfactory item fit and measurement invariance. Implications for future research are discussed.*

*Keywords:* students with special educational needs; testing accommodations; reading competence; large-scale assessment

[1] *Corresponding author*: Anna Südkamp, Emil-Figge-Str. 50, 44227 Dortmund, Germany, Phone: +49 231 755 6570, Fax: +49 231 755 6572, E-Mail: anna.suedkamp@tu-dortmund.de DOI: http://dx.doi.org/10.14786/flr.v3i2.130

# 1. Introduction

Large-scale assessments generally aim at drawing inferences about individuals' knowledge, competencies, and skills (Popham, 2000). Today, educational assessments play an important role as they inform students, parents, educators, policymakers, and the public about the effectiveness of educational services (Pellegrino, Chudowsky, & Glaser, 2F001). Using results from large-scale assessments, researchers can study factors influencing the acquisition and development of competencies and derive strategies on the improvement of educational systems. Often, assessments are meant to serve even more ambitious purposes such as supporting student learning (Chudowsky & Pellegrino, 2003). Assessing students' domain-specific competencies (e.g., reading competence, mathematical competence) is a key aspect of most large-scale assessments today (Weinert, 2001).

In this study, we focus on the assessment of competencies of students with special educational needs (SEN) in large-scale assessments. While national large-scale assessments like the National Assessment of Educational Progress (NAEP) in the United States and international assessments like the Programme for International Student Assessment (PISA) have established sophisticated methods for the assessment of students without SEN, testing students with SEN has proven to be challenging. In order to inform strategies for the assessment of students with SEN, we evaluate whether and if so, how students with SEN may be tested reliably and comparably to general education students. For this purpose, students with and without SEN were tested with accommodated and non-accommodated test versions. On the level of the single items, we carefully checked the reliability and comparability of the test scores obtained with the different test versions as reliability and comparability are necessary prerequisites for drawing meaningful inferences from large-scale assessments.

## 1.1 Assessing Reading Competence of Students With SEN

Large-scale assessments usually aim at describing the abilities of students within a country across the whole spectrum of the educational system or even across countries. This also includes students with SEN. In our notion, students with SEN include all students who are provided with special educational services due to a physical or mental impairment. In Germany, special schools are established for students with SEN. The special school system—in turn—is highly differentiated itself. There are special schools for students with special educational needs in learning, visual impairments, hearing disability/impairment, specific language/speech impairments, physical handicaps/disabilities, severe intellectual impairment/disability, emotional and behavioral difficulties, comprehensive SEN, and students with health impairment.

So far, comparatively little is known about the educational careers of students with SEN and their development of competencies across the life span (Heydrich, Weinert, Nusser, Artelt, & Carstensen, 2013; Ysseldyke et al., 1998). However, there is evidence that for students with SEN, reading problems pose one of the greatest barriers to success in school (Kavale & Reece, 1992; Swanson, 1999). Learning to read is a tedious process requiring psycholinguistic, perceptual, cognitive, and social skills (Gee, 2004). Beyond the basic acquisition of the alphabet system (i.e., letter-sound correspondence and spelling patterns), reading expertise implies phonological processing and decoding skills, linguistic knowledge (vocabulary, grammar), and text comprehension skills (Durkin, 1993; Verhoeven & van Leeuwe, 2008). According to Kintsch (2007), text comprehension can be seen as a combination of text-based processes that integrate previous knowledge to a mental representation of the text. It is thus a form of cognitive construction in which the individual takes an active role. Text comprehension entails deep-level problem-solving processes that enable readers to construct meaning from text and derives from the intentional interaction between reader and text (Duke & Pearson, 2002; Durkin, 1993).

On average, students with SEN show lower reading performance in large-scale assessments than students without SEN (Thurlow, 2010; Thurlow, Bremer, & Albus, 2008; Ysseldyke et al., 1998). For example, for the NAEP 1998 reading assessment in grades 4 and 8, Lutkus, Mazzeo, Zhang, and Jerry (2004) report lower average scale scores for students with SEN compared to students without SEN. Within the German KESS study (Bos et al., 2009) reading competence of seventh graders in special schools was compared to the reading competence of fourth graders in general education settings. Results demonstrated that fourth grade primary school students outperformed students with SEN in seventh grade in reading competence, the difference being about one third of a standard deviation. Drawing on data from a three-year longitudinal study, Wu et al. (2012) found that, compared to their general education peers, students receiving special educational services were more likely to score below the 10th percentile for several years in a row. In light of these findings, different reasons for the low performance of students with SEN have been discussed (Abedi et al., 2011). First, some students with SEN have difficulties related to the comprehension of text (e.g., lack of knowledge of common text structures, restricted language competencies, inappropriate use of background knowledge while reading; Gersten, Fuchs, Williams, & Baker, 2001). Reading problems of students with SEN in upper elementary and middle school are likely to be complex and heterogeneous resulting, for example, from a lack of phonological processing and decoding skills, a lack of linguistic knowledge (vocabulary, grammar), and a lack of text comprehension skills, or from a combination of problems in these areas. Second, lower performance could be attributed to a lack of opportunities to learn and to low teacher expectations (Woodcock & Vialle, 2011). Third, there could be barriers for students with disabilities in large-scale assessments that lead to unfair testing conditions (Pitoniak & Royer, 2001). According to Thurlow (2010), a combination of all these factors is likely. Taking the norm of test fairness seriously, large- scale studies try to ensure that students with disabilities will not be confronted with unfair testing conditions. That is why testing accommodations are often employed for students with SEN.

### 1.2 Providing Students with SEN With Testing Accommodations

The provision of testing accommodations for individuals with disabilities is a highly controversial issue in the assessment literature (Pitoniak & Royer, 2001; Sireci, Scarpati, & Li, 2005). Generally, testing accommodations are defined as changes in test administration that are meant to reduce construct-irrelevant difficulty associated with students' disability-related impediments to performance. According to the Standards for Educational and Psychological Testing, accommodations comprise "any action taken in response to a determination that an individual's disability requires a departure from established testing protocol. Depending on circumstances, such accommodation may include modification of test administration processes or modification of test content" (American Educational Research Association, 1999, p. 110). Note that some authors differentiate between *accommodations* and *modifications*: While accommodations are not meant to change the nature of the construct being measured, modifications result in a change in the test and equally affect all students taking it (Hollenbeck, Tindal, & Almond, 1998; Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998). In this article, however, we use the definition of the Standards for Educational and Psychological Testing.

Due to the many types of disabilities, various accommodations have been provided when testing students with SEN. Accommodations include, for example, modification of presentation format—including the use of braille or large-print booklets for visually-impaired examinees and the use of written or signed test directions for hearing-impaired examinees—and modification of timing, including extended testing time or frequent breaks (Koretz & Barton, 2003). In the 1998 NAEP reading assessment, a sample of students with varying disabilities and students with limited English proficiency were assigned to the following accommodations based on their individual needs: one-on-one testing, small-group testing, extended time, oral reading of directions, signing of directions, use of magnifying equipment, and use of an aide for transcribing responses.

Changes in the test bear the possibility that they alter the construct measured. If accommodated tests for students with SEN measure a different construct than the standard test for general education students, the competence scores between the two student groups are not comparable. Thus, it is utterly important to test whether test accommodations result in reliable and comparable competence measures (Borsboom, 2006;

Millsap, 2011). Lutkus et al. (2004) address the issue of whether the NAEP reading construct remains comparable for accommodated versus non-accommodated students by analyzing differential item functioning (DIF). DIF exists when subjects with the same trait level have a different probability of endorsing an item. Only very few items were found to have statistically significant DIF for the focal group (accommodated students) versus the reference group (non-accommodated students), which indicated measurement invariance across subgroups being assessed with different tests. In contrast, Koretz (1997) did find indications of DIF as 13 of 22 common items showed strong DIF when comparing item difficulty for students with SEN tested with accommodations and students without SEN tested under standard conditions using data from the Kentucky Instructional Results Information System assessment. In PISA 2012, samples of students with SEN were also tested with accommodated test versions (a shortened test version of the standard test and a test version including easier items). Here, the results on the psychometric properties of the accommodated test versions are still to be published (Müller, Sälzer, Mang, & Prenzel, 2014). In sum, the results concerning the use of testing accommodations are inconsistent. A major concern remains that in some cases accommodations may alter the test to the extent that accommodated and non-accommodated tests are no longer comparable (Abedi et al., 2011; Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, 2001; Cormier, Altman, Shyyan, & Thurlow, 2010).

However, one drawback of the studies by Lutkus et al. (2004) and Koretz (1997) is that in the analyses, different accommodations are not distinguished although different accommodations may have different effects. Another disadvantage is that students with SEN are often compared to students without SEN at the same grade level. Here, students with SEN and students without SEN differ not only in terms of their SEN status but also in terms of their expected achievement level. The study by Yovanoff and Tindal (2007) is one of the rare studies using an alternative comparison group where students with SEN in grade 3 are compared to students without SEN in grade 2.

Another issue is that comparisons usually involve students without SEN receiving the standard test and students with SEN receiving the accommodated test versions. By doing this, possible DIF may be due to both, testing accommodations and problems of testing students with SEN. In order to disentangle the appropriateness of test accommodations from the testability problems of students with SEN, the effects of test accommodations should separately be tested in a group of students without SEN. In the same vein, Pitoniak and Royer (2001) identify three major challenges for research on testing accommodations: variability in examinees, variability in accommodations, and small sample sizes (also see Geisinger, 1994). In the present study, we approach these challenges by focusing on students with special educational needs in learning (SEN-L), by focusing on specific accommodations appropriate for students with SEN-L, and by using a study design that incorporates a group of low-achieving students without SEN for evaluating the appropriateness of testing accommodations.

## 1.3 Testing Students With Special Educational Needs in Learning (SEN-L)

While providing students with physical, hearing, and visual impairments with testing accommodations is rather accepted, Pitoniak and Royer (2001) stress the importance of studying the effects of testing accommodations on test validity (or comparability), especially when testing students with learning disabilities. In this study, we focus on students with SEN-L in Germany, who comprise all students, who are provided with special educational services due to a general learning disability[1]. In Germany, students are assigned to the SEN-L group when their learning, academic achievement, and/or learning behavior are impaired (KMK, 2012) and when students cognitive abilities are below normal range (Grünke, 2004). In contrast to students with SEN-L, students with (specific) learning disabilities (e.g., a reading disorder) are not necessarily impaired in their general cognitive abilities. In Germany, the decision of whether a student

[1] As for the term "learning disabilities", the term SEN-L is not clearly defined. Note that we refer to a heterogeneous group of students with multifaceted etiology.

has special educational needs in learning is based on a diagnostic procedure and made collaboratively by parents, teachers, consultants, and school administrations. About 78% of the SEN-L students in Germany (KMK, 2012) do not attend regular schools but attend special schools with specific programs and trainings tailored to those who are unable to follow school lessons and subject matter in regular classes.

In fact, students with SEN-L compose the largest group of students with special educational needs in Germany (KMK, 2012). Comparably, students with learning disabilities compose the largest group of students with disabilities in the Unites States (Cortiella & Horowitz, 2014; US Department of Education, 2013). Our assumption is that the acceptance of testing accommodations for students with SEN-L is low, because the disabilities of students with SEN-L (e.g., information processing restrictions) are very likely to interfere with the construct that is to be measured (e.g., reading literacy). In turn, respective testing accommodations are likely to be construct-relevant. There are two test accommodations typically implemented for students with SEN-L: extended test time and "out-of-level" testing. Extended test time is usually implemented in order to compensate for information-processing restrictions in students with SEN-L. In his review on the appropriateness of extended time accommodations for students with SEN—including students with SEN-L among others—Lovett (2010) identified two studies with a serious amount of differentially functioning items, while DIF was negligible in one other study. A prominent hypothesis regarding extended test time is the *differential boost hypothesis* (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000), which states that students with SEN benefit more from extended time than students without SEN. In their review on test accommodations for students with SEN including 14 studies on extended time, Sireci et al. (2005) conclude that students with SEN as well as students without SEN benefit from extended test time. In only one of the reviewed studies students with SEN benefited more from extended test time than students without SEN.

Another common method is to provide students with SEN-L with an out-of-level test, which was originally meant for testing younger children (Thurlow, Elliott, & Ysseldyke, 1999). Similarly, alternate assessments that test lower-level reading and mathematical skills or skills that are precursory to reading and numerical literacy can be applied (Zebehazy, Zigmond, & Zimmerman, 2012). Both methods aim at avoiding undue frustrations for students with SEN-L and at improving the accuracy of measurement. Critics of out-of-level and alternate assessments argue that students with SEN-L are faced with low expectations due to the assessment and are prevented from taking the standard tests, and thus consider the assessments to be inappropriate for accountability assessment. Nevertheless, Thurlow et al. (1999) consider out-of-level testing a good opportunity to test students with SEN, if one can make sure that a common scale across different disparate grade levels is available. Such a common scale may be achieved by using methods of Item Response Theory (IRT), given that the items measure the same construct. When scaling Oregon's early reading alternate assessment onto the first general statewide benchmark reading assessment in grade 3, Yovanoff and Tindal (2007) identified good psychometric properties of the alternate assessment and no severe DIF between students with SEN (grade 3) and students without SEN (grade 2). However, data that support either the use or nonuse of out-of-level testing or alternate assessments is still rare (Minnema, Thurlow, Bielinski, & Scott, 2000; see the study by Zebehazy et al., 2012, which focuses on visually impaired students, for an exception).

## 2. Research Questions

As prior research has shown, testing competencies of students with SEN-L represents a challenge for large-scale assessments (Thurlow, 2010). Assessing competencies of students with SEN-L with tests that have been developed for students without SEN-L may fail to result in satisfying item fit measures and may be associated with differential item functioning, which impedes the opportunity to compare the competence scores of students with and without SEN. The present study aims to evaluate different strategies of testing students with SEN-L. Generally, we address the question of whether and how satisfying item fit measures and measurement invariant test scores can be obtained for students with SEN-L in large-scale-assessments. We evaluate whether standard tests developed for students without SEN-L and testing accommodations for students with SEN-L result in reliable and comparable measures of reading competence. If a reliable and comparable measurement of reading competence can be achieved, substantial research on the competence level, predictors of reading competence and competence development, as well as group differences may be investigated.

In this study, two major research questions are addressed: First, we investigate whether a reduction in test difficulty and a reduction of the number of items lead to test results comparable to students tested without accommodations. We approach these questions by testing students in general education, for whom reliable and valid competence scores can be obtained using a standard reading test. As the accommodated test versions are targeted towards a lower competence level, we did not use the whole group of students in general education, but focused on the subgroup of low-achieving students. Secondly, we explore whether these accommodations are suitable for testing students with SEN-L.

## 3. Method

### 3.1 Sample and Design

We collected data within the German National Educational Panel Study (NEPS). The NEPS is a national, large-scale longitudinal multicohort study that investigates the development of competencies across the lifespan (Blossfeld & von Maurice, 2011; Blossfeld, von Maurice, & Schneider, 2011). The study aims at providing high-quality, user-friendly data on competence development and educationally relevant processes for an international scientific community (Barkow et al., 2011).

Between 2009 and 2012, six representative start cohorts (Aßmann et al., 2011) were sampled, including about 60,000 individuals from early childhood to adulthood. Specific target groups include migrants (Kristen et al., 2011) and students with SEN-L (Heydrich et al., 2013). All participants are accompanied on their individual educational pathways through a collection of data on competencies (Weinert et al., 2011), learning environments (Bäumer, Preis, Roßbach, Stecher, & Klieme, 2011), educational decisions (Stocké, Blossfeld, Hoenig, & Sixt, 2011), and educational returns (Gross, Jobst, Jungbauer-Gans, & Schwarze, 2011). Following the principles of universal design (Dolan & Hall, 2001; Thompson, Johnstone, Anderson, & Miller, 2005), the NEPS aims at providing a basis for fair and equitable measures of competencies for all individuals.

In the present study, we used data from three different studies of students in fifth grade. These studies comprise a) a representative sample of general education students (main sample), b) a sample of students with SEN-L, and c) a group of students in the lowest academic track (LAT). The response rate in these studies was 55%, 45%, and 63%, respectively. In the main sample there were $N$ = 5,208 general education students, including $N$ = 700 students in the lowest academic track (see Aßmann, Steinhauer, & Zinn, 2012, for more information on the NEPS main sample). On average, these students were $M_{\text{age}} = 10.95$ ($SD_{\text{age}} = .53$) years old and 48.3% were female (0.7% had a missing response on age, 0.2% had a missing response on gender). About 24.1% of the students reported that they spoke a language other than German at home. The sample of students with SEN-L draws on a feasibility study with $N$ = 433 students who were recruited at special schools for children with SEN-L in Germany. Students in this sample were $M_{\text{age}} = 11.41$ ($SD_{\text{age}} = .63$) years old and 43.3% were female (0.7% had a missing response on gender). In this sample, about 30.1% of the students reported that they spoke a language other than German at home.

In this feasibility study, we applied two accommodated test versions that aimed at a) reducing the difficulty of the test and b) reducing the test length (and thereby increasing the testing time per item). In order to discern whether test items do not function properly because the accommodations change the test construct or whether students with SEN-L still have problems with the test, we implemented a group of low achieving students without SEN. This group consisted of a separate sample of $N = 490$ students enrolled in the lowest academic track, or *Hauptschule*. Students in this sample were $M_{\text{age}} = 11.28$ ($SD_{\text{age}} = .63$) years old and 48.4% were female. About 29.8% of the students in the LAT spoke a language other than German at home.

Focusing on this sample, we evaluated whether the accommodated test versions yield reliable test scores and whether they assess the same construct as the standard reading test. Students without SEN were tested as for this group it has already been shown that reliable and valid competence assessment can be obtained using the standard reading test. Thus, we could investigate the impact of the testing accommodations and disentangled testing problems resulting from badly-constructed accommodated test versions and testing problems resulting from the assessment of students with SEN-L. We restricted our sample to students in the lowest academic track, because the accommodated test versions were targeted towards a lower competence level. For students in general education in higher academic tracks, the test accommodations would be too easy and, as a consequence of such low test targeting, could result in aberrant response patterns (due to motivation problems) as well as in low item discriminations (due to the low variability in item responses). Implementing this group of low-achieving students allowed us to investigate whether the accommodated test versions generally result in reliable and comparable measures of competence.

All students were tested in the middle of fifth grade in November and December 2010. Data were collected by the International Association for the Evaluation of Educational Achievement (IEA) Data Processing and Research Center (DPC). Students participated in the study voluntarily, so student and parental consent was necessary. Each student who participated in the study received 5 euros.

## 3.2 Measures and Procedures

Within all three samples, reading literacy as well as mathematical competence was assessed. The orientation towards the functionality and everyday relevance of the competencies studied is one central aspect of the NEPS framework for the assessment of competencies. It draws on the concept of literacy in international comparative studies with a focus on enabling participation in society (see OECD, 1999). In this study, we focus on the assessment of reading literacy. Within the NEPS, the reading competence assessment focuses on text comprehension. All reading tests are developed based on a framework for the assessment of reading competence (Gehrer, Zimmermann, Artelt, & Weinert, 2013). This framework has been developed based on theoretical and pragmatic considerations that take earlier concepts and studies of reading competence within large-scale assessments into account. The most important dimensions within the framework are text types, cognitive requirements, and task formats. Concerning text types, texts with commenting, information, literacy-aesthetic, instruction, and advertising functions are included. In turn, cognitive requirements range from finding information in the text, drawing text-related conclusions, and reflecting and assessing. Across all age groups, the items in the test are either simple multiple choice (MC) items, complex MC items, or matching items. Complex multiple-choice (CMC) items present a common stimulus followed by a number of MC questions with two response options each. Matching (MA) items consist of a common stimulus followed by a number of statements, which require assigning a list of response options to these statements (see Gehrer, Zimmermann, Artelt, & Weinert (2012) for a full description of the framework including information on text types, cognitive requirements, item formats, and example items).

### 3.2.1 Standard reading test

The standard reading test was designed for students enrolled in the regular school system. It was developed based on the conceptual framework sketched above. Students were asked to read five different texts and answer questions focusing on the content of these texts (Gehrer, Zimmermann, Artelt, & Weinert, 2013). The test for students in fifth grade included a text about a continent (information function), a recipe (instruction function), an invitation (advertising function), a critical statement on a societal topic (commenting function), and a fictive story about a famous character (literacy-aesthetic function). In the analysis of the standard reading test, 56 items were included; however, subtasks of complex MC and matching items were treated as single items. So when combined, there were 33 questions in the standard reading test, which students had to complete within 30 minutes. For testing general education students, the test has shown good psychometric properties (Pohl, Haberkorn, Hardt, & Wiegand, 2012).

### 3.2.2 Reading test with accommodations

Based on the standard reading test, two accommodated test versions were administered in this study. As mentioned above typical testing accommodations for students with SEN-L include extended testing time and "out of level" testing. Within the NEPS, time for testing a domain-specific or domain-general competence is limited to 30 minutes. Under this restriction, we decided to develop one accommodated test version by reducing test length (*reduced test*), resulting in an increased test time per item. One text and its respective nine items, plus an additional 10 hard items were removed. The text on the societal topic and the items were removed, because the items showed to be comparatively difficult in prior item analyses in samples of general education students. In order to facilitate scaling of the different test versions on a same scale, an anchor item design (e.g., Kolen & Brennan, 2004) was used for linking the different test versions. For this design a sufficient number of items need to be the same in all test forms. Therefore, in the reduced test four texts and 37 items remained the same as in the standard reading test and functioned as anchor items in this design. We refer to the term "anchor item" when an item is the same in the standard test and in the accommodated test versions. As a result of reducing the length of the test, one text function was left out in the reduced test version (the commenting function). Still, the anchor items represented all three cognitive requirements. Note that while this accommodation mainly served to reduce test length, it also reduced item difficulty.

We decided to develop a second accommodated test version (*easy test*) that mainly aimed at reducing the difficulty of the standard test. Therefore, three texts and their respective 37 items from the standard reading test were removed (the text about the continent, the critical statement on a societal topic, and the fictive story about a famous character). These texts and its respective items were replaced with three texts and 23 items that had been developed for younger children in grade 3—including a text on the human body (information function), a short story about a family (literacy-aesthetic function), and an invitation (advertising function). This procedure can be considered as some sort of "out-of-level" testing. However, two texts remained the same as in the standard reading test as we used an anchor item design. Based on prior item analysis in samples of general education students in grade 5, five especially difficult items were eliminated from these texts. This procedure resulted in 12 overlapping items in the standard reading test and the easy test version. These items were used as anchor items in this design. In sum, the easy test version included 35 items.

Overall, 5,208 general education students including 700 students from the lowest academic track were tested with the standard reading test. Students with SEN-L took the standard reading test ($N$ = 176), the reduced test ($N$ = 173), or the easy test ($N$ = 84) by random assignment. The additional sample of $N$ = 490 students from the lowest academic track was randomly assigned to the reduced test ($N$ = 332) and the easy test ($N$ = 158). Note that the standard reading test was not administered to this sample of students in the lowest academic track. For investigating the appropriateness of the standard reading test for students in the

LAT, the subsample of the main sample of general education students attending schools of the lowest academic track were used (*N* = 700).

In order to control for fatigue and acquaintance effects, the order of the different tests was rotated within the booklet in almost all test versions. The standard test and the easy test were administered either before or after a mathematics test. For the analyses, due to sample size issues, the test order was ignored and the different conditions were analyzed together. Due to sample size limitations, there was no rotation of the position of the reduced test; the reduced test was only administered before the mathematics test. For the comparison of estimated item difficulties with general education students, data from these students refer to the same position within the booklet as data of students with SEN-L or the students in the LAT. So no bias is to be expected from test position.

## 4. Analyses

### 4.1 The Model

We scaled the data within the framework of Item Response Theory (IRT). In accordance with the scaling procedure for competence data in the NEPS (Pohl & Carstensen, 2012; 2013), we used a Rasch model (Rasch, 1960) estimated in ConQuest (Wu, Adams, Wilson, & Haldane, 2007). In this model a unidimensional measurement model with equal loadings across items is proposed. Various fit indices are available that describe the psychometric properties of the tests.

As described above, the reading test included complex MC and matching items. These items consisted of a set of subtasks that were aggregated to a polytomous variable in the final scaling model in the NEPS. When aggregating the responses on the subtasks to a single polytomous super-item, we lose information on the single subtasks. Since in this study we were interested in the fit of the items, we treated the subtasks of complex MC and matching items as single dichotomous items in the analyses. As such, we could not account for possible local item dependence within each set of subtasks. We applied the Rasch model to every test version (standard test, reduced test, easy test) and sample (students with SEN-L, students in the LAT).

### 4.2 Item Fit

In order to investigate whether the standard test and the accommodated reading tests reliably measured reading competence, we evaluated different fit measures. These included the weighted mean square (WMNSQ; Wright & Masters, 1982), item discrimination, point-biserial correlation of the distractors with the total score and the empirically approximated item characteristic curve (ICC). All of these measures provide information on how well the items fit a unidimensional Rasch model.

As Wu (1997) showed, fit statistics depend on the sample size. The larger the sample size, the smaller the WMNSQ and the greater the t-value. Thus, since the group of students with SEN-L differs in sample size from the group of students in the LAT, we considered different evaluation criteria for the interpretation of the WMNSQ. In this study, we report item discrimination, which describes the point-biserial correlation of the item with the total score (i.e., relative number of correct responses on the total number of valid responses). A well-fitting item should have a high positive correlation—that is, subjects with a high ability should score higher on the item than subjects with a low ability. For an easier interpretation, we report the discrimination not only in absolute values, but classify the item fit regarding the discrimination into acceptable item fit (discrimination > .2), slight misfit (discrimination between .1 and .2) and strong misfit (discrimination < .1). Furthermore, point-biserial correlations of incorrect response options and the total score are evaluated. The correlations of the incorrect responses with the total score allow for a thorough investigation of the performance of the distractors. A good item fit would imply a negative or zero correlation of the distractor with the total score. Distractors with a high positive correlation may indicate an ambiguity in relation to the correct response. Finally, empirically approximated item characteristic curves (ICC) were considered. These describe whether the number of correct responses corresponds to the theoretical implied response probability at each competence level.

### 4.3 Measurement Invariance

Reading scores of students with SEN-L versus students in general education can only be compared when the tests are measurement invariant—that is, when there is no differential item functioning (DIF). Measurement invariance is furthermore a necessary assumption for linking the different test forms. When measurement invariance holds—and thus there is no DIF—the probability of endorsing an item is the same for students with SEN-L and those without SEN-L who have the same ability. The presence of DIF is an indication that the respective reading test measures a different reading construct for both target groups, and thus that the reading scores between the target groups may not  be compared.

We tested DIF for each test version (standard, reduced, easy) and each target group (students with SEN-L, students in the LAT) by comparing the estimated item difficulties in the respective test version and target group to the estimated item difficulty of the same items for students in the main sample of the NEPS. Students with SEN-L as well as students in the LAT were, thus, compared to general education students in the main sample. There is one exemption: The group of students in the LAT was not tested with the standard reading test. In order to estimate DIF for that group on the standard test, we used the data of the students in the lowest academic track of the main sample of general education students. For this, we separated the main sample into students in the lowest academic track and students in other tracks and compared the estimated item difficulty between both groups.

We estimated DIF in a multi-facet IRT model, estimating separate item difficulties for general education students and for the respective target group. In line with the benchmarks chosen in the NEPS (Pohl & Carstensen, 2012), we considered absolute differences in item difficulties greater than 0.6 to be noticeable and absolute differences greater than 1 to be strong DIF. Note that these benchmarks serve here as an orientation for interpretation. To get a thorough picture, we also report the absolute DIF value. Also note that while in the standard test DIF may be investigated for all items, DIF in the reduced test and the easy test may only be investigated for the anchor items. In the reduced test and the easy test there are anchor items that allow linking of the different test versions. As described above, there are 37 anchor items in the reduced test and 12 anchor items in the easy test.

## 5. Results

In the following we will first represent the occurrence of missing values in each test form. Then we will present item fit for the different test forms and samples, followed by a further investigation of reasons for item misfit. In a next step, results on the comparability of test scores are presented. The results on item fit and measurement invariance are then considered together for evaluating the appropriateness of the different test forms for assessing competencies of students with SEN-L.

## 5.1 Missing Responses

Table 1 depicts the mean of the relative amount of different kinds of missing responses for each of the target groups and test versions. Similar to the main study (Pohl et al., 2012), there is a large number of missing responses—on average, up to 19% of the items are missing. The amount of missing responses is larger in the students with SEN-L group than in the group of students in the LAT for all test versions and all types of missing responses.

Comparing the different test versions, the lowest number of omitted items is found in the easy test version. This is probably due to the fact that the easy test version contains many easy items and that omission of items is related to the difficulty of the item (see, e.g., Pohl, Gräfe, & Rose, 2014). The lowest number of not reached items is found in the reduced test version. Thus, the reduction of texts and items to work on within the given assessment time does increase the number of items reached. The lowest number of invalid missing responses occurs in the reduced test version. This is likely because the reduced test version contains fewer matching items; this is the item format with the largest number of invalid responses (Pohl et al., 2012).

Table 1

*Averages of the Relative Frequency of Missing Responses*

| Type of missing response | Test | SEN-L | LAT |
|---|---|---|---|
| | | *M* | *M* |
| Omitted | standard | 6.72 | 4.78 |
| | reduced | 5.20 | 2.70 |
| | easy | 2.01 | 0.92 |
| Not reached | standard | 10.46 | 9.45 |
| | reduced | 3.90 | 1.04 |
| | easy | 5.63 | 3.44 |
| Invalid | standard | 1.13 | 0.44 |
| | reduced | 0.48 | 0.18 |
| | easy | 1.59 | 0.18 |
| Total number of missing responses | standard | 18.31 | 14.67 |
| | reduced | 9.58 | 3.92 |
| | easy | 9.22 | 4.53 |

*Note.* SEN-L = Special educational needs in learning; LAT = Lowest academic track.

## 5.2 Item Fit

### *5.2.1 Standard test*

First we analyzed item fit for the standard reading test for students with SEN-L and students in the lowest academic track. Overall, item discrimination is relatively small for students with SEN-L. The mean item discrimination is .25 (it is .34 in the lowest academic track). Four items show a slight misfit (discrimination between .1 and .2) and 10 items a strong misfit (discrimination less than .1). In the lowest academic track, there is only one item with a strong misfit and nine items with a slight misfit.

Evaluation of further fit measures for students with SEN-L confirms these results. Table 2 depicts the number of misfitting items for the WMNSQ, ICC, and point-biserial correlations. Summarizing these results, there is a large amount of items in the standard test that do not fit. EAP-Reliability of competence scores for students in the lowest academic track is sufficiently high (Rel = 0.823), while it is considerably lower for students with SEN-L (Rel = 0.652).We can conclude that students with SEN-L may not be tested appropriately with the standard reading test. In contrast, fit indices in the lowest academic track indicate a relatively good item fit that is comparable to the fit found in the main sample of general education students (see Pohl et al., 2012 for the results in the main study). The results indicate that the test is appropriate not only for the main sample including students attending higher academic tracks but also for low-performing students.

Table 2

*Number of Items With Misfit Indicated by Weighted Mean Square (WMNSQ), Item Characteristic Curve (ICC), and Point-Biserial Correlations*

| Fit measure | Test | SEN-L | LAT |
|---|---|---|---|
| WMNSQ | Standard | 7 | 7 |
| | Reduced | 2 | 1 |
| | Easy | 1 | 3 |
| ICC | Standard | 15 | 9 |
| | Reduced | 17 | 2 |
| | Easy | 12 | 1 |
| Point-biserial correlations | Standard | 21 | 3 |
| | Reduced | 14 | 1 |
| | Easy | 5 | 0 |

*Note.* SEN-L = Special educational needs in learning; LAT = Lowest academic track.

### 5.2.2 *Reduced test*

The item discriminations of the items in the reduced test version indicate a better item fit for students in the LAT than for students with SEN-L. For both target groups, the reduced test shows better item fit indices than the standard test version. For students with SEN-L, there are six items with a slight misfit (discrimination between .1 and .2) and five items with a strong misfit (discrimination below .1). Note that—not necessarily—the items showing misfit in the standard reading test, also show low discriminations in the reduced test. This may indicate that problems with testing of students with SEN-L do not necessarily lay in the specificity of the items, but may reflect other aspects of testing. The mean item discrimination is .28. In contrast, for students in the LAT the mean item discrimination is .47 and there is only one item with a slight misfit and one item with a strong misfit. Note that the item with the strong misfit was also problematic in the main sample. Evaluation of the WMNSQ, the ICCs, as well as of the point-biserial correlations of the responses (see Table 2) corroborates these findings. The results show that the items in the reduced test version have a good item fit for students in the LAT. They have, however, an insufficient fit in the students with SEN-L group. Nevertheless, the item fit in the students with SEN-L group is better for the reduced test than for the standard test. As in the standard test, EAP-reliability was sufficiently high for students in the lowest academic track (Rel = 0.850) but it was not sufficient for students with SEN-L (Rel = 0.525).

### 5.2.3 *Easy test*

The items in the easy test fit the data for both target groups better than the standard test. For students with SEN-L there are only four items with a slight misfit and three items with a strong misfit. The mean item discrimination for students with SEN-L is .30, while it is .46 for the students in the LAT. In the students of the LAT group, there is no item with an unsatisfactory discrimination. Also the other fit measures evaluated (see Table 2) show that the items in the easy test version fit the model in the group of students in the LAT but show some misfit in the students with SEN-L group. The EAP-reliability for students in the lowest academic track was high (Rel = 0.877), while it was not satisfactory for students with SEN-L (Rel = 0.600) Compared to the other two test versions, the easy test version shows the best model fit for students with SEN-L.

## 5.3 Investigation of Item Misfit

We further investigated the occurrence of item misfit based on test characteristics. We did not find any systematic relationship between item misfit and the different dimensions of the conceptual framework of the reading test (text function, cognitive requirements, and item format). However, we did find a relationship between item misfit and item difficulty.

### *5.3.1 Standard test*

The correlation of the item difficulty estimated in the main sample—thus being independent of the measurement model in the SEN-L group—and item discrimination within the students with SEN-L group is -.492. The more difficult an item, the lower is the discrimination. This may be an indication of disadvantageous test targeting—that is, inappropriate item difficulties for this target group. The items in the standard test are too difficult for students with SEN-L (mean item difficulty with the mean of the reading ability set to zero = 0.58 logits), while item difficulties match the abilities of the students of the lowest academic track well and are in fact rather easy (mean item difficulty = -0.41 logits). Here, the correlation between item difficulty estimated in the main sample and item discrimination for students in the lowest academic track is -.324. Note that since the measurement model of the standard test in the lowest academic track was estimated based on a subsample of the main sample, estimated item difficulty is not independent of the estimated item discrimination in the sample of students of the lowest academic track in the main sample.

### *5.3.2 Reduced test*

In the group of students in the LAT item fit of the reduced test is not substantively correlated with item difficulty (*cor* = -0.06) and is considerably negatively correlated in the students with SEN-L group (*cor* = -.43). Within students in the LAT, there is no relationship between item difficulty and item misfit, while in the students with SEN-L group, items with high difficulty show larger item misfit. This may also be a result of the small variance in item discrimination in the group of students in the LAT for this test version. Test targeting shows that the reduced test is still too difficult for students with SEN-L (mean item difficulty = 0.43 logits) but too easy for students in the lower academic track of general education (mean item difficulty = -1.03 logits).

### *5.3.3 Easy test*

Since most of the items in the easy test are not part of the standard test, we did not compute correlations between item difficulty and item fit. However, we did investigate test targeting. In test targeting, the easy test version is also too easy for students in the LAT (mean item difficulty = -0.99 logits) and too hard for students with SEN-L (mean item difficulty = 0.61 logits). Note that the easy test version is even more difficult than the reduced test version.

## 5.4 Measurement Invariance

### *5.4.1 Standard test*

Table 3 shows the absolute differences in estimated item difficulties first, between general education students and students with SEN-L and second, between students in the lowest academic track and students in other tracks of the main sample taking the standard test version. For students with SEN-L, negative values in the table indicate a higher item difficulty compared to general education students while positive values indicate lower item difficulty. For students in the lowest academic track, negative values indicate a higher item difficulty for these students compared to students in other tracks in the main sample and positive values indicate a lower item difficulty.

Table 3

*Differential Item Functioning (DIF) in the Different Test Versions and Student Groups*

| | | Differential Item Functioning | | | | | |
|---|---|---|---|---|---|---|---|
| Item | Difficulty | SEN-L | | | LAT | | |
| | | Standard | Reduced | Easy | Standard | Reduced | Easy |
| REG50110 | -1.909 | -1.010 | -0.942 | | -0.304 | -0.256 | |
| REG50121 | -2.814 | -1.678 | -1.200 | | -0.320 | 0.246 | |
| REG50122 | -2.063 | -0.926 | -0.800 | | -0.276 | -0.360 | |
| REG50123 | -2.078 | -0.444 | -0.848 | | -0.222 | -0.072 | |
| REG50124 | -2.236 | -0.510 | -0.930 | | -0.140 | -0.246 | |
| REG50125 | -2.202 | -1.018 | -0.752 | | -0.260 | -0.442 | |
| REG50126 | -1.793 | -0.652 | -0.512 | | 0.018 | 0.858 | |
| REG50127 | -2.173 | -0.714 | -1.234 | | -0.414 | -0.362 | |
| REG50130 | -0.805 | -0.850 | -0.090 | | -0.068 | -0.106 | |
| REG50140 | -0.148 | -0.382 | -0.288 | | -0.132 | -0.024 | |
| REG50150 | 0.874 | -0.400 | | | 0.200 | | |
| REG50161 | 0.542 | -1.688 | -1.388 | | -0.536 | -0.302 | |
| REG50162 | 0.149 | -1.020 | -0.130 | | -0.310 | -0.686 | |
| REG50163 | 0.035 | -0.348 | 0.422 | | -0.342 | -0.330 | |
| REG50164 | -0.076 | -1.320 | -0.774 | | -0.466 | -0.116 | |
| REG50165 | 0.048 | -0.302 | -0.042 | | -0.428 | -0.368 | |
| REG50170 | 2.351 | 0.570 | | | -0.294 | | |
| REG50210 | -1.411 | -1.054 | -0.566 | -0.564 | -0.352 | -0.574 | -0,304 |
| REG50220 | 1.436 | 1.200 | 1.602 | 1.360 | 0.576 | 0.414 | 0,490 |
| REG50230 | -1.187 | -0.926 | -0.814 | -0.850 | -0.148 | 0.006 | -0.148 |
| REG50240 | 0.050 | -0.082 | 0.146 | 0.232 | -0.094 | -0.044 | -0.226 |
| REG50250 | 0.667 | 0.164 | 0.344 | -0.096 | 0.134 | 0.170 | -0.018 |
| REG50261 | -1.352 | -0.318 | | | -0.204 | | |
| REG50262 | 1.924 | 0.580 | | | -0.038 | | |
| REG50263 | 2.159 | 0.172 | | | 0.088 | | |
| REG50264 | 2.167 | -0.290 | | | 0.188 | | |
| REG50265 | 2.195 | 0.724 | | | 0.180 | | |
| REG50266 | 2.221 | 1.016 | | | 0.116 | | |
| REG50310 | -0.867 | -0.824 | -1.254 | -0.756 | -0.318 | -0.142 | -0.444 |
| REG50320 | -1.425 | -0.982 | -0.870 | -0.798 | -0.464 | -0.196 | -0.066 |
| REG50330 | -1.185 | -1.654 | -1.632 | -1.020 | -0.440 | -0.106 | -0.154 |
| REG50340 | -0.158 | -0.570 | 0.378 | 0.026 | -0.186 | 0.078 | 0.030 |
| REG50350 | 0.838 | 0.082 | 0.310 | 0.420 | 0.102 | 0.028 | 0.222 |
| REG50360 | -0.887 | -0.844 | -0.324 | -0.324 | -0.274 | -0.062 | -0.130 |

(continued)

| Item | Difficulty | SEN-L | | | LAT | | |
|---|---|---|---|---|---|---|---|
| | | Standard | Reduced | Easy | Standard | Reduced | Easy |
| REG50370 | 0.140 | -0.256 | 0.318 | -0.058 | 0.020 | 0.288 | -0.100 |
| REG50410 | 0.885 | 0.370 | | | 0.206 | | |
| REG50421 | -0.481 | 0.042 | | | 0.342 | | |
| REG50422 | -0.225 | 0.380 | | | 0.666 | | |
| REG50423 | 0.243 | 1.268 | | | 0.536 | | |
| REG50430 | 2.371 | 0.772 | | | 0.080 | | |
| REG50452 | 0.531 | 1.586 | | | 0.590 | | |
| REG50440 | 1.922 | 1.264 | | | 0.374 | | |
| REG50451 | 0.183 | 1.526 | | | 0.716 | | |
| REG50460 | 1.356 | 0.436 | | | 0.100 | | |
| REG50510 | -0.898 | -0.532 | -0.618 | | -0.594 | 0.044 | |
| REG50521 | -0.313 | -0.052 | 0.878 | | -0.366 | -0.058 | |
| REG50522 | -0.635 | -0.156 | 0.966 | | -0.080 | 0.416 | |
| REG50523 | -0.004 | 0.366 | 1.066 | | 0.214 | 0.250 | |
| REG50524 | -0.634 | 0.256 | 0.872 | | -0.318 | 0.704 | |
| REG50530 | 1.487 | 0.770 | | | 0.206 | | |
| REG50540 | 0.064 | -0.262 | 0.748 | | -0.428 | -0.096 | |
| REG50551 | -0.035 | 0.064 | -0.756 | | 0.030 | -0.090 | |
| REG50552 | 1.135 | 0.188 | 1.214 | | -0.184 | -0.108 | |
| REG50553 | 0.385 | 0.210 | 0.540 | | -0.452 | 0.214 | |
| REG50560 | 1.125 | 0.716 | 0.938 | | 0.624 | 0.666 | |
| REG50570 | 0.515 | -0.334 | 0.392 | | -0.330 | 0.206 | |

Note. SEN-L = Special educational needs in learning; LAT = Lowest academic track.

The results clearly show measurement invariance for students in the lowest academic track and large differences in estimated item difficulties for students with SEN-L. For students in the lowest academic track, of the 56 items there is no item with strong DIF (absolute difference in item difficulties greater than 1) and only three items with slight DIF (absolute difference in item difficulties between 0.6 and 1). For students with SEN-L there are 12 items with slight DIF and 14 items with strong DIF. The results indicate that measurement invariance holds for students in the lowest academic track but that the test measures a different construct for the group of students with SEN-L compared to general education students. Thus, reading test scores for students with SEN-L are not comparable to test scores for general education students.

### *5.4.2 Reduced test*

Table 3 also shows DIF for the accommodated test versions. In the reduced test, for students with SEN-L, 15 out of 38 items have slight DIF and eight items have strong DIF. Only 15 items show no considerable DIF. Thus, the measurement of reading competence with the reduced test is different from that of general education students with the standard test. This does, however, not seem to be a result of the test accommodation. Within the group of students in the LAT measurement invariance holds as only three items show slight DIF. The results indicate that for students with SEN-L the measurement model, and thus, the measured construct, is different from that of students in general education.

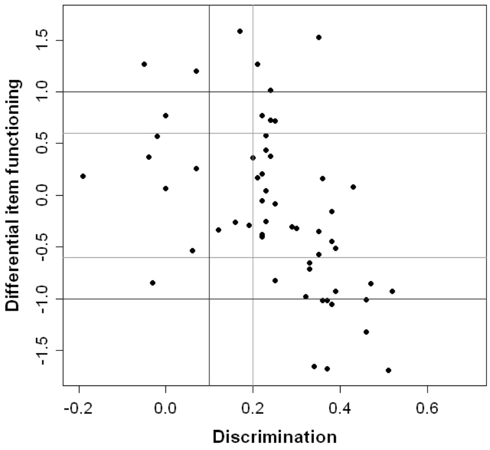
### *5.4.3 Easy test*

In the easy test, for students with SEN-L three out of twelve anchor items show noticeable DIF and two items show strong DIF. There are only seven items with no noticeable DIF. In contrast, in the LAT group there is no noteworthy DIF in the easy test and only four items show slight DIF in the reduced test.

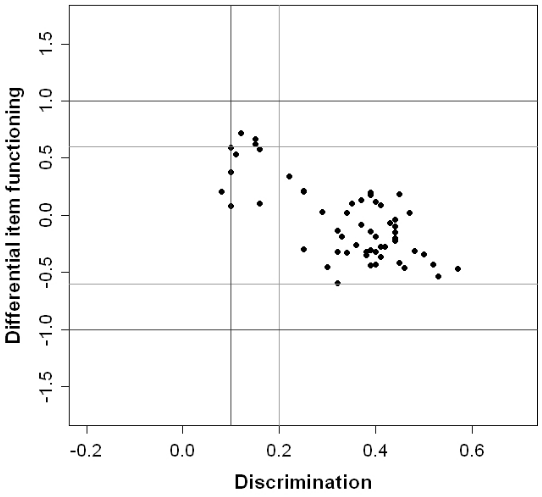
While measurement invariance may be assumed for the students in the LAT, it does not hold for students with SEN-L. Again, differences in the measurement model do not seem to be induced by the test accommodation, but rather reflect a specific testing problem of students with SEN-L.

## 5.5 Item Fit and Measurement Invariance

Considering both criteria—item fit and measurement invariance—how many items with good psychometric properties are left within the different groups and test versions? Is it possible to construct a test out of well-fitting items? Figure 1 shows the discrimination and DIF of the items in the standard test version for students with SEN-L (a) and for students of the lowest academic track (b). The grey lines give the rules of thumb for the evaluation of the items. Items within discrimination > .2 and absolute DIF < 0.6 have no noticeable misfit or DIF. Items within .2 > discrimination > .1 and 0.6 < absolute DIF < 1 have noticeable but not considerable misfit and/or DIF. Items with discrimination < .1 and absolute DIF > 1 have considerable misfit and/or DIF. These items should not be used for testing. Figure 1a) shows that a considerable amount of items do not meet the fit and DIF criteria in the SEN-L group. Only 22 out of 56 items show good fit and DIF indices. Thirteen items show a slight misfit in at least one of the two criteria and 21 items exceed at least one of the criteria for a strong misfit or large DIF. There are obviously not many items left that meet the criteria of a good test. For students of the lowest academic track (Figure 1b), there is only one item with a slight misfit in either of the two criteria and seven items with a strong deviation from at least one of the two criteria. Thus, there are 48 items that meet the criteria of a good test in the lowest academic track group of the main sample.



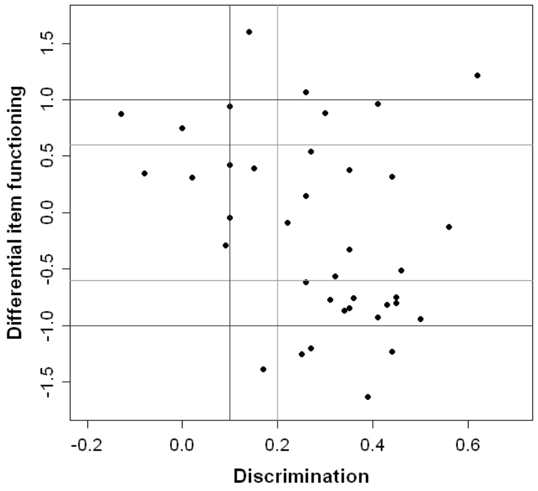
a) Students with SEN-L



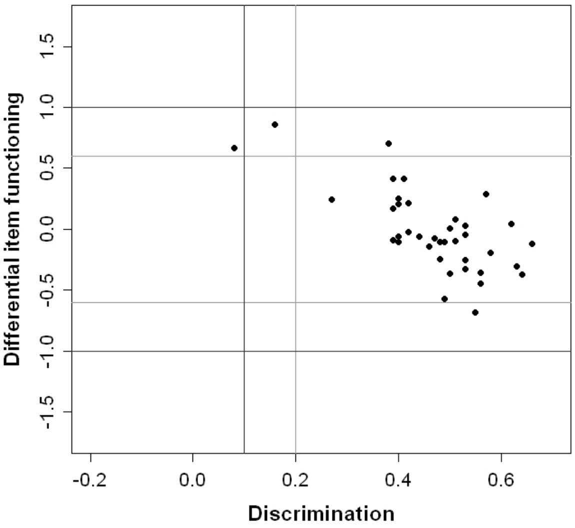
b) Students in the lowest academic track of the main sample

*Figure 1*. Discrimination and differential item functioning of the items in the regular test. SEN-L = Special educational needs in learning.

In the reduced test (see Figure 2a), for students with SEN-L, 13 out of 38 items show a strong misfit and/or DIF, 16 items show a slight deviation from at least one of the two criteria and only nine items are suitable for testing considering both criteria. There are a high number of items that may not be used on a test. Again, in the LAT group the items fit both criteria very well (Figure 2b). Only one out of 38 items needs to be excluded due to strong misfit or DIF, and only three items show a slight misfit and/or DIF. Thirty-four items meet the criteria of fit and measurement invariance. The low DIF values in the LAT group provide evidence in support of the argument that reducing the test length (i.e., increasing the testing time per text and item) does not threaten the comparability of the results. Thus, reducing test length may be an appropriate accommodation. However, this accommodation is not sufficient to reliably and comparably measure reading competence for students with SEN-L.



a) Students with SEN-L



b) Students in the lowest academic track

*Figure 2*. Discrimination and differential item functioning of the items in the reduced test. SEN-L = Special educational needs in learning.

Since there are only 12 items in the easy test that may be tested for DIF, we refrained from plotting the different evaluation criteria for this test version. It may, however, be concluded that from the 12 items, there are four with a slight misfit or DIF and two with a strong one. Only six of the 12 anchor items meet the criteria of fit and DIF. Since linking may only be done using 12 items, losing six items due to fit and DIF problems raises questions as to the appropriateness of this accommodated test version for the group of students with SEN-L. As a comparison, in the LAT group there is only one of these 12 items with a slight misfit and one with a strong misfit. The results in the LAT group are an indication that reducing the difficulty of the test does result in reliable and comparable reading competence measures. However, this test accommodation is not appropriate enough for assessing students with SEN-L.

## 6. Discussion

The present research dealt with the question of how competencies of students with SEN-L may be assessed reliably and comparably to general education students. We assessed the reading competence of students with SEN-L using a standard reading test, a reduced reading test, and an easy reading test. We used a group of low-achieving students without SEN to test whether the test accommodations alter the measured construct. The results showed that all three reading test versions are suitable for a reliable and comparable measurement of reading competence in students without SEN. Reducing both test length and item difficulty resulted in reliable measures that are comparable to those of a standard test for general education students. For students with SEN-L, the accommodated test versions considerably reduced the amount of missing values. They did not, however, show a satisfactory item fit and measurement invariance. Although the testing accommodations increase item fit and measurement invariance for students with SEN-L as compared to using a standard reading test, there are still many items unsuitable for a reliable and comparable assessment of reading competence in students with SEN-L. Thus, the competence scores assessed by the tests in this study are neither suitable for a substantive interpretation of the competence level of students with

SEN-L, nor may they be used for a valid comparison of competence levels between students with SEN-L and students in general education.

Concerning the testing accommodations implemented in this study, the *reduced test* primarily aimed at compensating for information-processing restrictions in students with SEN-L (e.g., for slow processing speed) while the *easy test* primarily aimed at adapting the test to a reduced competence level in reading (by reducing test difficulty in general) thereby improving the accuracy of measurement and avoiding undue frustrations for students with SEN-L. Since we showed—within the group of students in the LAT—that the items in the accommodated test versions have a good fit, we may conclude that the misfit in the group of SEN-L students is not due to badly constructed items or to the fact that the test versions changed the measured construct. Misfit of items in the SEN sample must be due to problems in testing this specific target group. Our analyses on test targeting showed that even the accommodated test versions are too difficult for students with SEN-L. Since item fit became better for accommodated versions, which were composed of easier items than the standard test, we hypothesize that a further reduction in item difficulty may help to improve testing of students with SEN-L. This hypothesis is corroborated by the negative correlation of item difficulty and discrimination. Still, both testing accommodations focus on general problems faced by students with SEN-L when reading (slow processing speed, reduced competence level in reading). In future research, it would be desirable to identify more specific reading problems of students with SEN-L that can be addressed in testing accommodations. Another explanation for item misfit in the sample of students with SEN-L may lay in the test-taking behavior (such as guessing or item omission, see Pohl, Südkamp, Hardt, Carstensen, & Weinert, 2015). It is also possible that differences in item fit between the students in the LAT and the students with SEN-L are due to differences in school curricula.

Comparing the three test versions—the standard test, the reduced test, and the easy test—in the LAT group, the accommodated test versions resulted in better competence measures than the standard test. For students with SEN-L, the easy test showed the best results regarding item fit, test targeting, and DIF. Since in the reading test, items are grouped to sets belonging to different texts, constructing a reading test from well-fitting and measurement invariant items is a difficult encounter. This is different in other competence domains of the NEPS that do not have such a strong testlet structure (see Weinert et al., 2011, for a description of the tests).

## 6.1 Strengths and Limitations

Studying the effects of testing accommodations not only in groups of students with SEN-L but also in groups of students in general education (here: low-performing students), is a promising approach to the identification of appropriate testing accommodations. In many previous studies, accommodated test versions were only applied to students with SEN. Thus, one could not disentangle whether low psychometric properties of accommodated tests and change of the measured construct were due to testing accommodations or testability problems of students with SEN. Using the LAT group allowed us to investigate whether the applied testing accommodations generally provide reliable and measurement invariant measures of reading competence. With the results in the group of LAT students, we ruled out the premise that misfit and measurement invariance for students with SEN-L is due to changes in the measured construct resulting from a reduction in test length or reduction in item difficulty. Considering the wide range of competence levels of students in general education, students in the LAT are the group of students without SEN being closest in competence level to students with SEN. Thus, the accommodated test versions—that are targeted towards students with SEN-L—will still be better targeted to students in the LAT than to all students in general education.

The study's strength also lies in the use of a sophisticated methodological approach and the evaluation of various measures of item fit in addition to differential item functioning. When using methods of IRT, other studies on the assessment of students with SEN mainly report DIF but leave out information on item fit in the sample of students with SEN (Abedi, Leon, & Kao, 2008; Bolt & Ysseldyke, 2008).

Considering the group of students with SEN-L, using data from a relatively large representative sample allows us to draw credible conclusions. However, our samples of students with SEN-L and students in the LAT group considerably differed in their size. There were about twice as many students in the LAT group compared to the students with SEN-L group. For some testing conditions the sample was comparatively small. For example, only 84 students with SEN-L were assessed with the reduced test version. Due to the large number of missing responses, there were items with just 52 valid responses. Fit and DIF measures may, as a consequence, be unreliable. We tried to account for this in the evaluation of the fit and DIF criteria.

One might also argue that the group of students with SEN-L is still a highly heterogeneous one, including, for example, students with different performance and ability profiles in the cognitive domain. Compared to prior research, however, the target population is rather homogeneous as students with SEN in areas other than learning (e.g., those with physical impairments) are precluded. Other studies investigated appropriateness of competence assessments on even more heterogeneous groups of students (e.g., Lutkus et al., 2004, including students with disabilities in general). Possible testing problems may, however, only occur for students with specific disabilities (e.g., for students with SEN-L, but not for students with visual impairments) or for specific testing accommodations. Analyzing the whole group of students with disabilities and running analyses across all types of testing accommodations may mask possible testing effects. In our study we focused on a specific group of students with SEN and analyzed different testing accommodations separately.

Item misfit and DIF do not need to be caused by all students with SEN-L, but only by a certain group of students. However, we did not account for interindividual differences within our samples in this study. In ongoing research, we (Pohl et al., 2015) use a person-based approach and try to empirically identify groups of students with SEN-L whose assessment is especially challenging. Here, we assume that individual student characteristics (e.g., individual test taking strategies, cognitive performance profiles) are related to testability[2].

### 6.2 Implications and Future Research

Incorporating easy instead of hard items in the test version (e.g., as done in the easy test version), is methodologically seen a form of adaptive testing. Adaptive testing is currently discussed in large-scale studies such as the NAEP (Xu, Sikali, Oranje, & Kulick, 2011), the Programme for International Student Assessment (PISA; Pearson, 2011), and the NEPS (Pohl, 2014). If better test targeting is one of the key issues for testing students with SEN-L, adaptive testing procedures for general education students may well be extended to include students with SEN-L. One way to systematically reduce difficulty in reading tests for students with SEN might be a reduction in grammatical and lexical complexity of texts and items (Abedi et al., 2011). In upcoming feasibility studies within the NEPS, seventh graders with SEN-L will be tested with a standard reading test that is reduced in grammatical and lexical complexity. In another feasibility study in grade 3 we will examine the effects of newly developed test instructions on students' test performance, missing values, and invalid answers, as well as on their motivation, and test anxiety.

There are numerous and manifold arguments for the inclusion of students with SEN in large-scale assessments. However, the issue of whether students with SEN-L may be assessed reliably and comparably in large-scale assessments —and if so how—remains to be an important and complex question. In our study, we aim to present a sophisticated design and a comprehensive methodological approach to these questions

[2] In the present study, differences in test taking in students with and without SEN-L might also be caused by differences in school curricula. This alternative hypothesis could be tested by comparing students with SEN-L attending general education and special schools. However, in Germany only few students with SEN-L attended general education schools at the time of data collection and these students often differ in individual as well as in social background characteristics from students attending special schools.

and to shed light on them. We think that the systematic identification of specific testing accommodations for groups of students with SEN is a promising approach.

## Keypoints

- So far, data on the acquisition and development of competencies of students with special educational needs in learning (SEN-L) are rare.
- Assessing competencies of students with special educational needs within large scale assessments is challenging.
- This study addresses the question of whether and how satisfying item fit measures and measurement invariant test scores can be obtained for students with SEN-L in large-scale-assessments.
- Testing accommodations may result in reliable and to the standard test comparable competence measures.
- The investigated testing accommodations helped to some extent to increase the testability of students with SEN-L.
- The systematic identification of further appropriate testing accommodations is a promising approach to the assessment of students with SEN-L.

## Acknowledgments

This paper uses data from the National Educational Panel Study (NEPS). From 2008 to 2013, NEPS data were collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

We especially thank Cordula Artelt, Claus H. Carstensen, Jana Heydrich, Lena Nusser, and Markus Messingschlager for their contribution to this study. Our thanks also go to the staff of the NEPS administration of surveys and to the methods group.

We would also like to thank the anonymous reviewers for their comments on earlier versions of the manuscript and Erika Fisher for copy editing services.

## References

Abedi, J., Leon, S., & Kao, J. (2008). *Examining differential item functioning in reading assessments for students with disabilities.* (CRESST Report 744). Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Leon, S., Kao, J., Bayley, R., Ewers, N., Herman, J., & Mundhenk, K. (2011). *Accessible reading assessments for students with disabilities: The role of cognitive, grammatical, lexical, and textual/visual features* (CRESST Report 785). Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., … Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. *Zeitschrift für Erziehungswissenschaft, 14*, 51-65. doi:10.1007/s11618-011-0181-8

Aßmann, C., Steinhauer, H. W., & Zinn, S. (2012). *Weighting the fifth and ninth grader cohort samples of the National Educational Panel Study, panel cohorts* (Technical Report). Bamberg, Germany: University of Bamberg National Educational Panel Study, Retrieved from https://www.nepsdata.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC3/1-0-0/SC3_SC4_1-0-0_Weighting_EN.pdf.

Bäumer, T., Preis, N., Roßbach, H.-G., Stecher, L., & Klieme, E. (2011). Education processes in life-course-specific learning environments. *Zeitschrift für Erziehungswissenschaft, 14*, 87-101. doi:10.1007/s11618-011-0183-6

Barkow, I., Leopold, T., Raab, M., Schiller, D., Wenzig, K., Blossfeld, H.-P., & Rittberger, M. (2011). RemoteNEPS: Data dissemination in a collaborative workspace. *Zeitschrift für Erziehungswissenschaft, 14*, 315-325. doi: 10.1007/s11618-011-0192-5

Bielinski, J., Thurlow, M. L., Ysseldyke, J. E., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items* (NCEO Technical Report 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Blossfeld, H.-P., & von Maurice, J. (2011). Education as a lifelong process. *Zeitschrift für Erziehungswissenschaft, 14*, 19-34. doi:10.1007/s11618-011-0179-2

Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft, 14*, 5-17. doi:10.1007/s11618-011-0178-3

Bolt, S. E., & Ysseldyke, J. (2008). Accommodating students with disabilities in large-scale testing: A comparison of differential item functioning (DIF) identified across disability types. *Journal of Psychoeducational Assessment, 26*, 121-138. doi:10.1177/0734282907307703

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*, 425-440. doi: 10.1007/s11336-006-1447-6

Bos, W., Bonsen, M., Gröhlich, C., Guill, K., May, P., Rau, A., et al. (2009). *KESS 7: Kompetenzen und Einstellungen von Schülerinnen und Schülern—Jahrgangsstufe 7* [KESS 7: Competencies and attitudes of students in grade 7]. Hamburg, Germany: Behörde für Bildung und Sport.

Chudowsky, N., & Pellegrino, J. (2003). Large-scale assessment that support student learning: What will it take? *Theory into Practice, 42*, 75-83. doi:10.1207/s15430421tip4201_10

Cormier, D. C., Altman, J., Shyyan, V., & Thurlow, M. L. (2010). *A summary of the research on the effects of test accommodations: 2007-2008* (Technical Report 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Cortiella, C., & Horowitz, S. H. (2014). *The state of learning disabilities: Facts, trends and emerging issues.* New York: National Center for Learning Disabilities.

Dolan, R. P., & Hall, T. E. (2001). Universal design for learning: Implications for large-scale assessment. *IDA Perspectives, 27*, 22-25.

Duke, N. K., & Pearson, P. D. (2002). Effective practices for developing reading comprehension. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (pp. 205–242). Newark, DE: International Reading Association.

Durkin, D. (1993). *Teaching them to read*. Boston, MA: Allyn and Bacon.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L., & Karns, K. M. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data. *School Psychology Review, 29*, 65–85.

Gee, J. P. (2004). Reading as situated language: A sociocognitive persepective. In R. B. Ruddell & N. J. Unrau (Eds.), *Theoretical models and processes of reading* (pp. 116-132). Newark: International Reading Association.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal of Educational Research Online, 5*, 50-79.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2012). *The assessment of reading competence (including sample items for grade 5 and 9)* [Scientific Use File 2012, Version 1.0.0.] Bamberg: University of Bamberg, National Educational Panel Study.

Geisinger, K. F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education, 7*, 121-140. doi:10.1207/s15324818ame0702_2

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of Educational Research, 71*, 279-320. doi:10.3102/00346543071002279

Gross, C., Jobst, A., Jungbauer-Gans, M., & Schwarze, J. (2011). Educational returns over the life course. *Zeitschrift für Erziehungswissenschaft, 14*, 139-153. doi:10.1007/s11618-011-0195-2

Grünke, M. (2004). *Lernbehinderung* [Learning Disabilities]. In Lauth, G., Grünke, M., & Brunstein, J. (Eds.). Interventionen bei Lernstörungen [Interventions to learning deficits](pp. 65-77). Göttingen: Hogrefe.

Heydrich, J., Weinert, S., Nusser, L., Artelt, C., & Carstensen, C. H. (2013). Including students with special educational needs into large-scale assessments of competencies: Challenges and approaches with the German National Educational Panel Study (NEPS). *Journal of Educational Research Online, 5,* 217-240.

Hollenbeck, K. Tindal, G. Almond, P. (1998). Teachers' knowledge of accommodations as a validity issue in high-stakes testing. *The Journal of Special Education, 32*, 175-183.

Kavale, K. A., & Reece, J. H. (1992). The character of learning disabilities. *Learning Disability Quarterly, 15*, 74-94. doi: http://dx.doi.org/10.2307/1511010

Kintsch, W. (2007). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [Standing Conference of the Ministers of Education and Cultural Affairs of Germany] (2012). *Sonderpädagogische Förderung in Schulen 2001–2010* [Special education in schools 2001–2010]. Retrieved from http://www.kmk.org/fileadmin/pdf/Statistik/KomStat/Dokumentation_SoPaeFoe_2010.pdf

Kolen M. J., & Brennan R. L. (2004). *Test equating, scaling, and linking*. New York, NY: Springer-Verlag.

Koretz, D. M. (1997). *The assessment of students with disabilities in Kentucky* (CSE Technical Report 431). Los Angeles, CA: CRESST/RAND Institute on Education and Training.

Koretz, D. M., & Barton, K. E. (2003). *Assessing students with disabilities: Issues and evidence* (CSE Technical Report 587). Los Angeles, CA: University of California, Center for the Study of Evaluation.

Kristen, C., Edele, A., Kalter, F., Kogan, I., Schulz, B., Stanat, P., & Will, G. (2011). The education of migrants and their children across the life course. *Zeitschrift für Erziehungswissenschaft, 14*, 121-137. doi:10.1007/s11618-011-0194-3

Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research, 80*, 611-638. doi:10.3102/0034654310364063

Lutkus, A. D., Mazzeo, J., Zhang, J., & Jerry, L. (2004). *Including special-needs students in the NAEP 1998 reading assessment part II: Results for students with disabilities and limited-English proficient students* (Research Report ETS-NAEP 04-R01). Princeton, NJ: ETS.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Minnema, J., Thurlow, M., Bielinski, J., & Scott, J. (2000). *Past and present understandings of out-of-level testing: A research synthesis.* (Out-of-Level Testing Project Report 1). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/OOLT1.html

Müller, K., Sälzer, C., Mang, J., & Prenzel, M. (2014, March). *Kompetenzen von Schülerinnen und Schüler mit besonderem Förderbedarf. Ergebnisse aus dem PISA 2012 Förderschul-Oversample* [Competencies of students with special educational needs. Results from the PISA 2012 oversample of special schools]. Paper presented at the Conference of the German Association for Empirical Educational Research, Frankfurt, Germany.

OECD – Organisation for Economic Co-Operation and Development. (1999). *Measuring student knowledge and skills: A new framework for assessment.* Paris, France: OECD.

Pearson (2011, October 7th). *Pearson to develop framework for OECD's PISA students assessment for 2015* [Pearson announcement]. Retrieved from http://www.pearson.com/news/2011/october/pearson-to-develop-frameworks-for-oecds-pisa-student-assessment-f.html?article=true

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, D. C.: National Academy Press.

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71*, 53-104. doi:10.3102/00346543071001053

Pohl, S. (2014). Longitudinal multi-stage testing. *Journal of Educational Measurement, 50*, 447-468. doi: 10.1111/jedm.12028

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report: Scaling the data of the competence test* (NEPS Working Paper No. 14). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Pohl, S., & Carstensen, C. H. (2013). Scaling the competence tests in the National Educational Panel Study—Many questions, some answers, and further challenges. *Journal for Educational Research Online, 5*, 189-216.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not reached items in competence tests - Evaluating approaches accounting for missing responses in IRT models. *Educational and Psychological Measurement*, 74, 423-452. doi: 10.1177/0013164413504926

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS technical report for reading—Scaling results of starting cohort 3 in fifth grade* (NEPS Working Paper No. 15). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Pohl, S., Südkamp, A., Hardt, K., Carstensen, C. H., & Weinert, S. (2015). *Testability and test-taking behavior of students with special educational needs in large-scale assessments*. Manuscript submitted for publication.

Popham, W. J. (2000). *Educational measurement.* Boston, MA: Allyn and Bacon.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche (Expanded Edition, Chicago, University of Chicago Press, 1980).

Ritchey, K. D., Silverman, R. D., Schatschneider, C., & Speece, D. L. (2015). Prediction and stability of reading problems in middle childhood. *Journal of Learning Disabilities, 48*, 298-309. doi:10.1177/0022219413498116

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*, 457-490. doi:10.3102/00346543075004457

Stocké, V., Blossfeld, H.-P., Hoenig, K., & Sixt, M. (2011). Social inequality and educational decisions in the life course. *Zeitschrift für Erziehungswissenschaft, 14*, 103-199. doi:10.1007/s11618-011-0193-4

Swanson. (1999). Reading research for students with LD: A meta-analysis of intervention outcomes. *Journal of Learning Disabilities, 32*, 504-532. doi:10.1177/002221949903200605

Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (NCEO Technical Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L. (2010). Steps toward creating fully accessible reading assessments. *Applied Measurement in Education, 23*, 121-131. doi:10.1080/08957341003673765

Thurlow, M. L., Bremer, C., & Albus, D. (2008). *Good news and bad news in disaggregated subgroup reporting to the public on 2005-2006 assessment results* (Technical Report 52). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M., Elliott, J., & Ysseldyke, J. (1999). *Out-of-level testing: Pros and cons* (Policy Directions No. 9). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/Policy9.htm

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64*, 439–450

U.S. Department of Education, National Center for Education Statistics. (2013). *Digest of Education Statistics, 2012* (NCES 2 014-015).

Verhoeven, L., & van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology, 22*, 407-423. doi:10.1002/acp.1414

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen, L. & H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45-66). Seattle: Hogrefe & Huber.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft, 14*, 67-86. doi:10.1007/s11618-011-0182-7

Woodcock, S., & Vialle, W. (2011). Are we exacerbating students' learning disabilities? An investigation of pre-service teachers' attributions of the educational outcomes of students with learning disabilities. *Annals of Dyslexia, 61*, 223-241. doi:10.1007/s11881-011-0058-9

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.

Wu, M. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalized item response models* (Unpublished doctoral dissertation). Melbourne, Australia: University of Melbourne.

Wu, M., Adams, R. J., Wilson, M., & Haldane, S. (2007). *Conquest 2.0*. [Computer Software] Camberwell, Australia: ACER Press.

Wu, Y.-C., Liu, K. K., Thurlow, M. L., Lazarus, S. S., Altman, J., & Christian, E. (2012). *Characteristics of low performing special education and non-special education students on large-scale assessments* (Technical Report 60). Minneapolis, MN: University of Minnesota, National Centre on Educational Outcomes.

Xu, X., Sikali, E., Oranje, A., & Kulick, E. (2011, April). *Multi-stage testing in educational survey assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), New Orleans, LA.

Yovanoff, P., & Tindal, G. (2007). Scaling early reading alternate assessments with statewide measures. *Exceptional Children, 73*, 184-201.

Ysseldyke, J. E., Thurlow, M. L., Langenfeld, K. L., Nelson, R. J., Teelucksingh, E., & Seyfarth, A. (1998). *Educational results for students with disabilities: What do the data tell us?* (Technical Report 23). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Zebehazy, K. T., Zigmond, N., & Zimmerman, G. J. (2012). Ability or access-ability: Differential item functioning of items on alternate performance-based assessment tests for students with visual impairments. *Journal of Visual Impairment & Blindness, 106*, 325-338.

## Table of Footnotes

[2] As for the term SEN-L, the term "learning disabilities" is not clearly defined. Note that we refer to a heterogeneous group of students with multifaceted etiology.

[3] In the present study, differences in test taking in students with and without SEN-L might also be caused by differences in school curricula. This alternative hypothesis could be tested by comparing

students with SEN-L attending general education and special schools. However, in Germany only few students with SEN-L attended general education schools at the time of data collection and these students often differ in individual as well as in social background characteristics from students attending special schools.