Including Students With Special Educational Needs in the Competence Assessment of the NEPS—Results on the Comparability of Test Scores in Reading

Anna Südkamp, Steffi Pohl, Jana Heydrich and Sabine Weinert

Abstract

Including students with special educational needs in learning (SEN-L) is one of the National Educational Panel Study's (NEPS) challenges. In this study, we address the question of whether the reading competence of students with SEN-L may be assessed reliably with the reading test designed for general-education students. In addition, we ask whether the test scores of students with SEN-L can be compared with the test scores of students without SEN-L. The reading competence of $N = 176$ students with SEN-L and $N = 5{,}208$ general-education students is assessed with the NEPS standard reading test for students in Grade 5. The results of test targeting and item fit reveal that the items of the NEPS standard reading test are rather difficult for students with SEN-L, while item discrimination is low for many items of the test. With respect to measurement invariance, a substantial number of items show differential item functioning, indicating that the standard reading test measures a different construct for students with and without SEN-L. Implications for further research are indicated in the discussion.

1 Introduction

Today, educational assessments play an important role in society as they inform students, parents, educators, policy-makers, and the public about the effectiveness of educational services (Pellegrino, Chudowsky, & Glaser, 2001). Using results from large-scale assessments, factors influencing the acquisition and development of competencies can be studied and strategies on the improvement of educational systems can be derived. Tests within large-scale assessments aim at a valid and reliable measurement of competencies while—at the same time—being both time- and cost-efficient. In order to assure objectivity, tests are usually administered under standard-

Hans-Peter Blossfeld et al. (eds.), *Methodological Issues of Longitudinal Surveys,*
DOI 10.1007/978-3-658-11994-2_27

ized conditions. Testing is a highly demanding situation from each of the different perspectives of test-administrators, test-takers, parents, and teachers (Guthrie, 2002). For example, Abrams, Pedulla, and Madaus (2003) report that teachers frequently feel pressured to raise test scores. At the same time, increased levels of anxiety, stress, and fatigue have been observed among students. When it comes to testing students with special educational needs (SEN), the challenges of testing seem to be even higher since there might be specific barriers in large-scale assessments for students with SEN (Bolt & Ysseldyke, 2008). For example, students with visual impairments may not be able to access printed material, and students with learning disabilities may not be acquainted with these kinds of tests. However, giving students with SEN the opportunity to participate in large-scale assessments is an issue of fairness and equality. It is also highly relevant for being able to address important practical as well as theoretical questions in research on the developmental and educational pathways for students with SEN. Therefore, efforts have been made to reduce barriers in large-scale assessments and to include more students with special educational needs. Assessing students' domain-specific competencies (e. g., reading or mathematical competence) is a key aspect of the National Educational Panel Study (NEPS;[1] Weinert et al., 2011). The NEPS is a national large-scale longitudinal 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 the international scientific community (Barkow et al., 2011). Between 2009 and 2012, six representative starting 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 special educational needs in learning (SEN-L; Heydrich, Weinert, Nusser, Artelt, & Carstensen, 2013). 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 order to empirically address the question of whether and how students with SEN-L can be tested fairly, the NEPS has set up a series of feasibility studies. These studies focus on the validity of competence assessments. For example, we study the effects of testing accommodations for students with SEN-L on the reliability and comparability of test scores. Testing accommodations are generally defined as changes in test administration that are meant to reduce construct-irrelevant difficulty associated with students' disability-related im-

1 This paper uses data from the National Educational Panel Study (NEPS). The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by the German Federal States. Our research is based on the dedicated work of professors and research assistants, particularly those within the NEPS. We especially wish to thank Cordula Artelt, Claus H. Carstensen, Lena Nusser, and Markus Messingschlager. Our thanks also go to the staff of the NEPS survey administration and to the methods group.

pediments to performance. To test for group-specific effects and the comparability of test results and in order to discern—if necessary—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 implement a control group of students from the lowest academic track, or *Hauptschule*. In addition, we gather in-depth background information on students with SEN-L in surveys of the students' parents, teachers, and school principals.

1.1 Inclusion of Students With Special Educational Needs (SEN) in Large-Scale Assessments

In Germany, the population of students with SEN comprises more than 485,400 individuals, which is around 6.4 % of the entire student population (KMK, 2012). The question at hand is whether students with SEN can be tested reliably with the same test instruments and under comparable testing conditions as students without SEN. In the literature and in the field, this question has often been answered in the negative for theoretical as well as empirical and practical reasons. Therefore, students with SEN are still not being extensively included in large-scale assessments. Schools that are solely attended by students with SEN are excluded at the very beginning of the sampling procedure in the Progress in International Reading Literacy Study (PIRLS) as well as in the Programme for International Student Assessment (PISA) (Joncas, 2007; OECD, 2012). Whether students with SEN who are enrolled in general-education schools are included in these studies is mainly decided upon by local school staff even though all studies provide material to alleviate the decision-making process. In PIRLS, students with SEN are included as far as they are able to participate under standard conditions; otherwise, they are excluded. Contrary to PIRLS, PISA provides an extra "one hour" booklet specifically designed for students with SEN that contains half of the items of the standard test (OECD, 2012, p. 29). Surprisingly, in spite of a thorough description of the test design, main national PISA reports on Germany do not even mention the use of this booklet (cf. OECD, 2010). Despite a lack of studies and research reports on students with SEN, there is evidence that reading problems pose one of the greatest barriers to success in school for students with SEN (Kavale & Reece, 1992; Swanson, 1999).

1.2 Reading Performance of Students With SEN

On average, students with SEN[2] show a lower reading performance in large-scale assessments in comparison with students without SEN (Thurlow, 2010; Thurlow, Bremer, & Albus, 2008; Ysseldyke et al., 1998). For the 1998 National Assessment of Educational Progress (NAEP) of reading in Grades 4 and 8, Lutkus, Mazzeo, Zhang, and Jerry (2004) report lower average scale scores for students with SEN in comparison with students without SEN. Within the German study "Kompetenzen und Einstellungen von Schülerinnen und Schülern" (Bos et al., 2009), reading competence of seventh graders in special schools was compared with the reading competence of fourth graders attending general-education settings. Results demonstrated that fourth-grade primary-school students outperformed students with SEN in the 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 students receiving special educational services were more likely to score below the 10th percentile for several years in a row compared with their general-education peers. 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., a lack of knowledge of common text structures, restricted language competencies, inappropriate use of background knowledge while reading; Gersten, Fuchs, Williams, & Baker, 2001). Second, lower performance could be attributed to low teacher expectations and/or to a lack of opportunities to learn (Woodcock & Vialle, 2011). Third, there could be barriers for students with disabilities that lead to unfair testing conditions in large-scale assessments (Pitoniak & Royer, 2001). According to Thurlow (2010), a combination of all these factors is likely. Taking the norm of test fairness seriously, the NEPS tries to ensure that students with SEN will not be confronted with unfair testing conditions.

1.3 Assessment of Students With SEN With Standard Reading Tests

Providing students with SEN with standard reading tests has the advantage that no changes to the standard test instrument are necessary. Whenever changing a test instrument, there is a risk that test scores will not be comparable between groups tested with the standard test and accommodated test versions. Research on testing accommodations (Lovett, 2010; Pitoniak & Royer, 2001) has shown that testing accommodations may significantly alter standard test instruments, leading to test scores that

2 Note that students with SEN comprise a highly heterogeneous group, including, for example, students with visual impairments, hearing disabilities/impairments, and emotional and behavioral difficulties.

are no longer comparable. Nevertheless, students with SEN are often tested with accommodated test versions in large-scale assessments for practical reasons (Bolt & Ysseldyke, 2008; Pitoniak & Royer, 2001). So far, only a few studies have addressed the question of whether this is actually necessary, that is, whether students with SEN can also be tested validly and reliably with standard reading tests. As an exception, Koretz and Hamiliton (2000; see also Koretz, 1997, for more detailed results) report that 19 % (Grade 4), 33 % (Grade 8), and 39 % (Grade 11) of students with SEN were tested without accommodations in the Kentucky Instructional Results Information System assessment. As data of students with SEN tested with and without accommodations were available, item difficulty, item discrimination, and differential item functioning (DIF) were analyzed. Unfortunately, not all results were reported (e. g., exemplifications of the target and reference group in DIF analyses are missing; DIF-values are not presented). Koretz (1997) concluded that item discriminations were comparable for students with and without SEN and that instances of DIF were few and generally minor for students with SEN who were tested without accommodations. In line with these results, Lutkus et al. (2004) did not identify any items with a strong indication of DIF for the 1999 NAEP reading assessment when comparing the results of students with disabilities tested without accommodations with the results of students without disabilities. Here, a split-sample design was implemented: Half of the sample of students with SEN were tested without accommodations, while the other half were tested with accommodations. In contrast, Bielinski et al. (2001) conclude—based on their item analyses including the root mean squared discrepancy and differential item functioning—that the reading test results of non-accommodated assessments of students with a primary disability in reading on the Missouri Assessment Program were not comparable with the results of other examinees. In summary, results on the comparability of test scores for students with and without SEN on standard reading assessments are mixed. Aside from differential item functioning, indicators of item fit are reported scarcely. Although testing accommodations are often used in the assessment of students with SEN in large-scale assessments, we consider it beneficial to first analyze whether testing students with SEN with standard test instruments is appropriate.

1.4 Research Questions

Taking the norm of test fairness seriously, the NEPS wants to ensure that students with disabilities are not confronted with barriers in the assessment. At the same time, we want to ensure reliable and valid measurements of competencies. While the need for specially developed test instruments is obvious for some students with special educational needs (e. g., providing visually-impaired students with tests in Braille), students with SEN-L can, in principle, be tested with standard-competence tests. However, psychometric problems (e. g., differential item functioning) might be expected. As students with SEN-L comprise the largest group of students with special educa-

tional needs (KMK, 2012; Koretz, 1997), the NEPS has decided to specifically focus on this group of students when setting up a series of feasibility studies in order to investigate whether and how valid competence measures can be obtained from students with SEN-L (Heydrich et al., 2013). In this chapter, we focus on the assessment of reading competence and report on an initial set of analyses based on the assessment of SEN-L students with the NEPS standard reading test (see Südkamp, Pohl, Hardt, Jordan, and Duchhardt (2015) for results on the NEPS assessment of mathematical competence). We address the question of whether students with SEN-L can be tested reliably with the NEPS standard reading test and whether the test results of students with SEN-L are comparable with those of general-education students.

2 Method

2.1 Sample and Design

The data of this study were collected within the NEPS. The study draws on two different samples within the NEPS: One concerns students with SEN-L, and the other concerns general-education students from the NEPS main sample. The sample of the feasibility studies comprised $N = 176$ students with SEN-L in fifth grade who were recruited at special schools for children with SEN-L in Germany. On average, these students were $M_{age} = 11.39$ ($SD_{age} = 0.65$) years old, and 46 % were female. In Germany, students are assigned to the group of students with special educational needs in learning, when their learning, academic achievement, and/or learning behavior is impaired (KMK, 2012). The decision of whether a student is in need of special education is usually made jointly by parents, teachers, consultants, and school administrations. About 78 % of the SEN-L students in Germany (KMK, 2012) do not attend regular schools but instead attend special schools with specific schooling programs and trainings tailored to those students who appear to be unable to follow school lessons and subject matter in regular classes. However, it is becoming more and more common to educate students with SEN-L at general-education schools as well. For the present study, students with SEN-L were exclusively drawn from special schools. As a reference group, the study draws on representative data from the NEPS main sample (Starting Cohort 3 in Grade 5; see Aßmann, Steinhauer, & Zinn, 2012, for more information on the sampling), which comprises $N = 5{,}208$ students in general-education schools ($M_{age} = 10.95$ years, $SD_{age} = 0.53$; 48.3 % female).

2.2 Measures and Procedures

Reading and mathematical competences were assessed within both samples. Within the NEPS, the assessment of reading competence focuses on text comprehension,

which is often conceived of as the essence of reading (Durkin, 1993; Verhoeven & Van Leeuwe, 2008). Across all ages, starting in Grade 5, individuals read five different texts and are asked questions focusing on the content of these texts (Gehrer, Zimmermann, Artelt, & Weinert, 2013). The standard reading test was designed for students enrolled in the regular school system. The test was developed based on a conceptual framework that comprises five different text functions or text types and three different cognitive requirements (finding information in a text, drawing text-related conclusions, reflecting and assessing content). The items in the test were either multiple-choice (MC) items, complex MC items, or matching items (see Gehrer, Zimmermann, Artelt, & Weinert, 2012, for a description of the item formats in the reading test). Overall, 56 items were included in the analyses; however, subtasks of complex MC and matching items were treated as single items. When combined, there were 33 questions in the standard reading test, which students had to complete within 30 minutes. The test shows good psychometric properties for testing general-education students (Pohl, Haberkorn, Hardt, & Wiegand, 2012).

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

2.3 Analyses

The model

We scaled the data within the framework of Item Response Theory (IRT). In accordance with the scaling procedure of competence data in NEPS (see Pohl & Carstensen, 2012), we used a Rasch model (Rasch, 1960) estimated in ConQuest (Wu, Adams, Wilson, & Haldane, 2007). As described above, the reading test also 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 we are interested in the fit of the items in this study, we treated the subtasks of complex MC and matching items as single dichotomous items in the analyses.[3]

3 Note that we do not account for possible local item dependence within each set of subtasks with this analysis strategy.

Test targeting
In order to investigate whether the standard reading test was adequately targeted to the ability of the students with SEN-L, we evaluated test targeting. To do this, the estimated item difficulties were depicted on the same scale as the ability estimates. A test is considered well targeted if the item difficulties cover the whole range of ability estimates and there is no superfluity of items at the lower (too easy) or upper (too hard) end of the ability distribution.

Measures of fit
In order to investigate whether the standard reading test reliably measured reading competence for students with SEN-L, we evaluated different fit measures. For this analysis, we focus on the item discrimination, which describes the correlation of the item with the total score. A well-fitting item should have a high positive correlation, that is, subjects with a high ability should be more likely to score high on the item than subjects with a low ability. We considered a discrimination below .2 as a slight misfit and discriminations smaller than .1 as a strong item misfit.

Differential Item functioning
The comparability of the reading score of SEN students with those of general-education students can only be assured when the tests are measurement invariant—that is, when there is no DIF. 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 general-education students 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 validly compared. We estimated DIF in a multi-group IRT model, estimating and comparing item difficulties for general-education students and students with SEN-L. 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.

3 Results

3.1 Test Targeting

Figure 1 depicts the estimated item difficulties and the ability estimates of students with SEN-L on the same scale (in logits). In this analysis, the mean of the student's ability is set to zero. Ability estimates greater than zero indicate an above-average reading ability, while ability estimates smaller than zero indicate a below-average reading ability. Test takers with an ability that corresponds to the difficulty of an item have a 50 % probability of solving the item. Items with a lower difficulty are solved

Figure 1 Test targeting of the standard test in the group of SEN-L students. Item difficulties are depicted on the right side, person ability on the left side. Each number represents an item.

Scale (in logits)	**Person ability**	**Item difficulty**
		12 25 26
		17
		40
2	XX	
	X	27
	XXXXX	
	XX	11 24
		15 28
	X	13
	X	53
	XX	44
	XX	56
	XXX	33 50
	XXXXXX	41
	X	
1	XXXXX	36
	XXXXXXX	22 31 32 35 55
	XXXXXX	14 51
	XXXXXXXX	16
	XXXXXXXX	10 19 54
	XXXXXXXXXXXXXX	21
	XXXXXXXXXXXXXXXXXXXX	9
	XXXXXXXXXXX	29 34
	XXXXXXXXXXXXXXXXXXXXXXX	52
	XXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXX	20 46
	XXXXXXXXXXXXXXXXXXXXXXXXXXX	18 45 48
0	XXXXXXXXXXXXXXXXXXXXXX	30 47
	XXXXXXXXXXXXXXXXXXXXXXXX	37
	XXXXXXXXXXXXXXXXXXXXXXX	38
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXX	1 49
	XXXXXXXXXXXXXXXXXXX	39
	XXXXXXXXXX	23 43
	XXXXXXXXXXXXXX	2 3 6 7
	XXXXXXXXXX	
	XXXXXXXXX	42
	XXXXXXX	
-1	XXXXXX	8
	XXXX	
	XX	4
	XXX	5
	XX	
	XX	
	XX	
	X	
	X	

Each "X" represents 0.4 cases

with a higher probability, while items with a higher difficulty are solved with a probability lower than 50 %. Figure 1 shows that the item difficulties cover the whole range of students' abilities. However, the test is rather difficult overall. The gross of items is targeted towards students with high reading abilities. As a consequence, students with SEN-L may be overstrained by the test. As a comparison, the test is a bit too easy for students in general education (Pohl et al., 2012).

3.2 Item Fit

In Figure 2, item discrimination is displayed for the standard reading test in the group of students with SEN-L. Overall, item discrimination is relatively small for students with SEN-L. The mean item discrimination is .25. Four items show a slight misfit (discrimination less than .2 and equal to or greater than .1), and 10 items display a strong misfit (discrimination less than .1). As a comparison, there is no item misfit in the group of general-education students with the exception of one item that was excluded from the analyses. The item discrimination levels for general-education students are all above .3 (Pohl et al., 2012).

We further investigated the occurrence of item misfit in the standard test by estimating the correlation of the item difficulty estimated on general-education students (which is thus independent of the measurement model for SEN-L students) and the discrimination in the sample of SEN-L students. Within the group of students with SEN-L, item difficulty and discrimination correlated to −.492. The more difficult an item, the lower the discrimination is. That misfit occurs due to a disadvantageous test targeting—that is, due to inappropriate item difficulties for this target group. The items in the standard test are too difficult for students with SEN-L (mean item difficulty = 0.58 logits[4]).

3.3 Measurement Invariance

Figure 3 shows the absolute differences in estimated item difficulties between general-education students and students with SEN-L who took the standard reading test. Positive values in the table indicate a higher item difficulty for general-education students as compared with students with SEN-L, while negative values indicate a lower item difficulty.

The results clearly show large differences in estimated item difficulties for students with SEN-L compared with general-education students. 12 out of 56 items have a slight DIF, and 14 items have a strong DIF. The results indicate that the test measures

4 Note that the mean of the reading ability is set to zero.

Figure 2 Discrimination of the items in the standard reading test for students with SEN-L

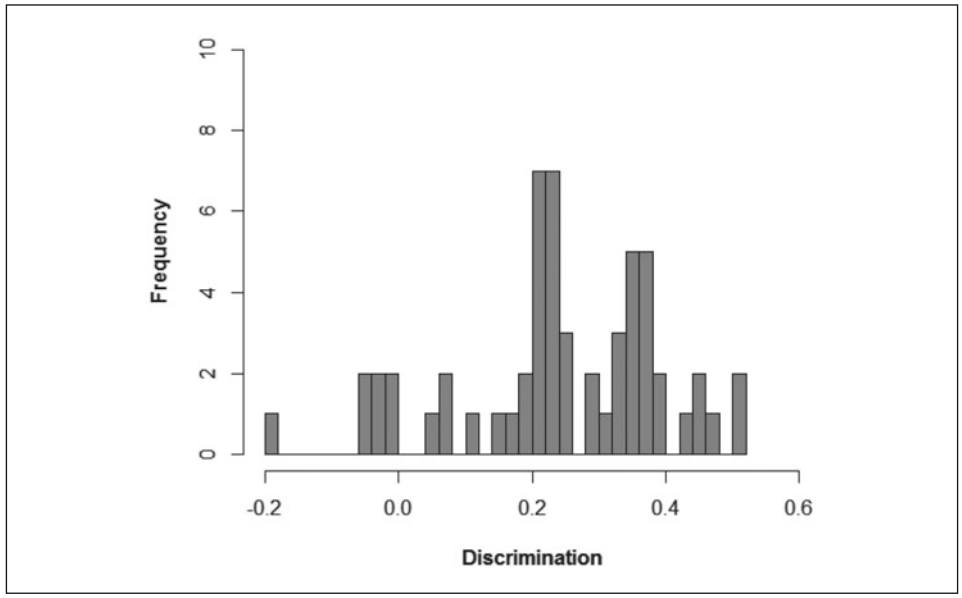
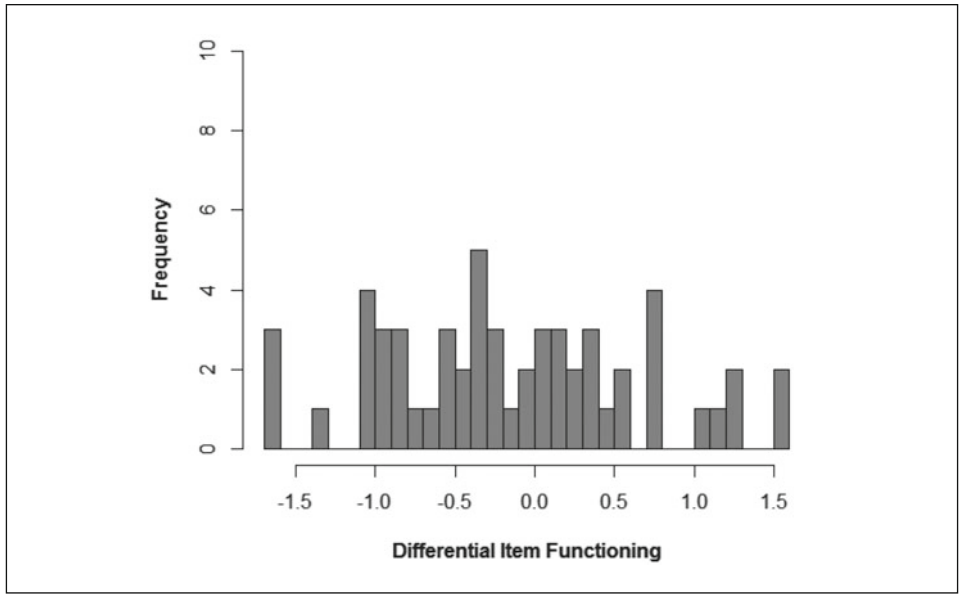


Figure 3 Differential item functioning of the items in the standard reading test. The graph depicts the differences in estimated item difficulties between students with SEN-L and general-education students



a different construct in the group of students with SEN-L as compared with general-education students. Reading-test scores for SEN-L students are thus not comparable with test scores for general-education students.

4 Discussion

The present study is part of a research program dealing with the question of how the competencies of students with SEN-L may be assessed reliably and comparably. In this chapter, we have addressed the question of whether the competencies of students with SEN-L in Grade 5 can be assessed reliably and comparably with the NEPS standard reading test. For this purpose, students with SEN-L were tested with the same test and under the same conditions as general-education students. As mentioned above, the standard reading test has shown good psychometric properties when testing high-achieving as well as low-achieving general-education students (Pohl et al., 2012).

The results on test targeting and item fit reveal that the items of the NEPS standard reading test are rather difficult for students with SEN-L. Item discrimination is low for many items of the test, showing that the items do not differentiate well between low-performing and high-performing students. With respect to measurement invariance, a substantial number of items show DIF, indicating that students with and without SEN-L cannot be measured on the same scale using the NEPS standard reading test.

With the present research, we contribute to the discussion of whether competencies of students with SEN may be assessed reliably and comparable by large-scale assessments. Our research overcomes problems of earlier studies on the assessment of students with SEN (see, e. g., Lovett, 2010). First, we concentrated our research on a specific group of students with SEN, namely students with learning disabilities. As such, we focus on a rather homogenous group of students and are able to disentangle whether the standard reading test is appropriate for a certain group of students with SEN.[5] In contrast, many other studies on students with SEN include students with various disabilities, which leads to samples that are even more heterogeneous. Second, our sample of students with SEN-L was tested with the age-appropriate standard reading test, regardless of students' disability status. Thus, we were able to study the psychometric quality of the test in a sample of students with SEN-L, while there was no selection of especially capable students with SEN-L. Third, the results of our analyses are based on a relatively large representative sample of students with SEN-L.

5 Please note that the group of students with SEN-L is still a heterogeneous one, including, for example, students with different performance and ability profiles in the cognitive domain. Compared with 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.

There is a complex research program in the NEPS dealing with the question of the testability of students with SEN-L within large-scale assessments. Within this program, the appropriateness of different aspects of testing is systematically investigated in order to identify appropriate testing conditions for students with SEN-L. The analyses reported in this chapter are the basis for further analyses. Südkamp, Pohl, and Weinert (2015), for example, investigated whether different testing accommodations result in reliable and comparable measures of reading competence. Testing accommodations include a reduction in test length as well as a reduction in the test's item difficulty. Further test accommodations draw on a reduction of grammatical and lexical complexity in the texts and items and on a specifically designed test-coaching phase prior to testing. Other research questions motivated by the present study are addressed by Pohl, Südkamp, Hardt, Carstensen, and Weinert (2015). These authors investigated whether there are differences in large-scale testability between students with SEN-L and how these differences are related to individual test-taking behavior.

References

Abedi, J., Leon, S., Kao, J., Bayley, R., Ewers, N., Herman, J., & Mundhenk, K. (2010). *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.

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice, 42,* 18–29. doi:10.1207/s15430421tip4201_4

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., … Blossfeld, H.-P. (2011). In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften. 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). Retrieved from https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC3/1-0-0/SC3_SC4_1-0-0_Weighting_EN.pdf

Barkow, I., Leopold, T., Raab, M., Schiller, D., Wenzig, K., Blossfeld, H.-P., & Rittberger, M. (2011). RemoteNEPS: Data dissemination in a collaborative workspace. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 315–325). Wiesbaden: VS Verlag für Sozialwissenschaften. 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 No. 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Blossfeld, H.-P., & von Maurice, J. (2011). Education as a lifelong process. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 19–34). Wiesbaden: VS Verlag für Sozialwissenschaften. 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. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 5–17). Wiesbaden: VS Verlag für Sozialwissenschaften. 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

Bos, W., Bonsen, M., & Gröhlich, C. (Hrsg). (2009). *KESS 7: Kompetenzen und Einstellungen von Schülerinnen und Schülern—Jahrgangsstufe 7* [KESS 7: Competencies and attitudes of students in grade 7]. Hamburg: Behörde für Bildung und Sport.

Dolan, R. P., & Hall, T. E. (2001). Universal design for learning: Implications for large-scale assessment. *IDA Perspectives, 27*(4), 22–25.

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

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.

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*(2), 50–79.

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

Guthrie, J. T. (2002). Preparing students for high-stakes test taking in reading. In A. E. Farstrup, & S. J. Samuels (Eds.), *What research has to say about reading instruction* (pp. 370–391). Newark: International Reading Association.

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 within the German National Educational Panel Study (NEPS). *Journal of Educational Research Online, 5*(2), 217–240.

Joncas, M. (2007). PIRLS 2006 sampling design. In M. O. Martin, I. V. S. Mullis, & A. M. Kennedy (Eds.), *PIRLS 2006 technical report* (pp. 35–48). Chestnut Hill, M.A: TIMSS & PIRLS International Study Center, Boston College.

Kavale, K. A., & Reece, J. H. (1992). The character of learning disabilities: An IOWA profile. *Learning Disability Quarterly, 15,* 74–94. doi: 10.2307/1511010

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] (Dokumentation No. 196). Berlin: Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

Koretz, D. M. (1997). *The assessment of students with disabilities in Kentucky* (CSE Technical Report No. 431). Los Angeles, CA: CRESST/RAND Institute on Education and Training.

Koretz, D. M., & Hamilton, L. (2000). Assessment of students with disabilities in Kentucky: Inclusion, student performance, and validity. *Educational Evaluation and Policy Analysis, 22,* 255–272. doi:10.3102/01623737022003255

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. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 121–137). Wiesbaden: VS Verlag für Sozialwissenschaften. 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.

OECD. (2010, December). *PISA 2009 results: What students know and can do: Student performance in reading, mathematics and science, 1.* Retrieved from http://dx.doi.org/10.1787/9789264091450-en

OECD. (2012, March). *PISA 2009 Technical Report.* Retrieved from http://dx.doi.org/10.1787/9789264167872-en

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

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: University of Bamberg, National Educational Panel Study.

Pohl, S., Südkamp, A., Hardt, K., Carstensen, C. H., & Weinert, S. (2015). *Testing students with special educational needs—Psychometric properties of test scores and associations with test taking behavior.* Manuscript submitted for publication.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

Südkamp, A., Pohl, S., Hardt, K., Duchhardt, C., & Jordan, A.-K. (2015). Kompetenzmessung in den Bereichen Lesen und Mathematik bei Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf [Competence assessment of students with special educational needs in the areas of reading and mathematics]. In P. Kuhl, P. Stanat, B. Lütje-Klose, C. Gresch, H. A. Pant, & M. Prenzel (Eds). *Inklusion von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen* (S. 243–272). Wiesbaden: VS Verlag für Sozialwissenschaften.

Südkamp, A., Pohl, S., & Weinert, S. (2015). Competence assessment of students with special educational needs—Identification of appropriate testing accommodations. *Frontline Learning Research, 3,* 1–25. doi:10.14786/flr.v3i2.130

Swanson, L. (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 No. 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 No. 52). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

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, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 67–86). Wiesbaden: VS Verlag für Sozialwissenschaften. 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

Wu, M., Adams, R. J., Wilson, M., & Haldane, S. (2007). *Conquest 2.0.* [Computer Software]. Camberwell: 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 No. 60). Minneapolis, MN: University of Minnesota, National Centre on Educational Outcomes.

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 No. 23). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

About the authors

A. Südkamp
Rehabilitation Psychology, TU Dortmund University,
Emil-Figge Str. 50, 44227 Dortmund, Germany.
e-mail: anna.suedkamp@tu-dortmund.de

S. Pohl
Methods and Evaluation/Quality Assurance, Free University Berlin,
Habelschwerdter Allee 45, 14195 Berlin, Germany.

J. Heydrich
Formerly University of Bamberg,
Wilhelmsplatz 3, 96045 Bamberg, Germany.

S. Weinert
Department of Psychology I: Developmental Psychology,
University of Bamberg,
Markusplatz 3, 96047 Bamberg, Germany.