

Secondary Publication



Gnambs, Timo; Nusser, Lena

Out-of-Level Cognitive Testing of Children with Special Educational Needs

Date of secondary publication: 22.02.2024

Version of Record (Published Version), Article

Persistent identifier: urn:nbn:de:bvb:473-irb-935999

Primary publication

Gnambs, Timo; Nusser, Lena (2024): „Out-of-Level Cognitive Testing of Children with Special Educational Needs“. In: European journal of psychological assessment : official organ of the European Association of Psychological Assessment, Vol. 40, Nr. 1, pp. 40-45, Göttingen ; Cambridge, Mass.: Hogrefe, doi: 10.1027/1015-5759/a000736.

Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holders.

This document is made available under a Creative Commons license.



The license information is available online:

<https://creativecommons.org/licenses/by/4.0/legalcode>



Out-of-Level Cognitive Testing of Children with Special Educational Needs

Timo Gnams  and Lena Nusser

Leibniz Institute for Educational Trajectories, Bamberg, Germany

Abstract: Children with special educational needs in the area of learning (SEN-L) have severe learning disabilities and often exhibit substantial cognitive impairments. Therefore, standard assessment instruments of basic cognitive abilities designed for regular school children are frequently too complex for them and, thus, unable to provide reliable proficiency estimates. The present study evaluated whether out-of-level testing with the German version of the *Cognitive Abilities Test* using test versions developed for younger age groups might suit the needs of these children. Therefore, $N = 511$ children with SEN-L and $N = 573$ low achieving children without SEN-L attending fifth grades in Germany were administered four tests measuring reasoning and verbal comprehension that were designed for fourth graders. The results showed that children with SEN-L exhibited significantly more missing responses than children without SEN-L. Moreover, three of the four tests were still too difficult for them. Importantly, no substantial differential response functioning was found for children with and without SEN-L. Thus, out-of-level testing might represent a feasible strategy to assess basic cognitive functioning in children with SEN-L. However, comparative interpretations would require additional norms or linked test versions that place results from out-of-level tests on a common metric.

Keywords: intelligence, reasoning, verbal comprehension, special educational needs, differential response functioning



Children who experience difficulties in learning, competence acquisition, and/or sustained attention are often assigned the label special educational needs in the area of learning (SEN-L; Grünke & Grosche, 2014; Lloyd et al., 2007). Their cognitive profiles are usually below average but rather heterogeneous and characterized by substantial inter- and intra-individual differences. In Germany, this label qualifies them for additional support in school and is assigned after a diagnostic process. Procedures of diagnosis are not standardized, and the selection of test instruments varies across federal states and school districts. However, valid and reliable diagnostics of cognitive skills require adequate instruments that differentiate at the respective proficiency level, whose item difficulties are not systematically biased for different groups and thus allow for comparison with others. Since most test developers do not acknowledge the unique challenges children with SEN-L may face during testing, the resulting instruments do not necessarily meet these criteria. Children with SEN-L have more difficulty comprehending test instructions (Nusser & Weinert, 2017)

and thus fail to work on items correctly or provide responses that comply with the instruction. They have less knowledge of and are less able to implement adequate strategies when responding to test items (Bosson et al., 2010). Moreover, children with SEN-L will most often show lower probabilities of solving items correctly compared to children without SEN-L (Gnams & Nusser, 2019), resulting in limited measurement accuracy. These group-specific challenges related to the content and the administration of a diagnostic measure could lead to invalid and incomparable data for these children. Studies examining well-established, comprehensive intelligence tests revealed unacceptable model fit and lack of measurement invariance for samples of adults with intellectual disabilities (e.g., MacLean et al., 2011). To address these issues, educational studies around the world implement accommodations to correct SEN-related barriers students might encounter. Such accommodations are manifold and can refer to the setting, the time, the format, or additional support (e.g., technology; Fuchs et al., 2005). Concerning the format, children may receive items developed for younger children, a so-called out-of-level test. In this case, the tests are expected to achieve a better fit between proficiency levels and item difficulty (e.g., Anderson et al., 2011; Minnema et al., 2001).

This brief report addresses the question of whether an out-of-level accommodation for four subtests of the German version of the *Cognitive Abilities Test* (Heller & Perleth, 2000) assessing reasoning and verbal comprehension represents an adequate accommodation for children with SEN-L attending fifth grade. Analyses will specifically investigate measurement accuracy, adequacy of item difficulty, and measurement invariance compared to children without SEN-L.

Methods

Sample and Procedure

As part of the German *National Educational Panel Study* (NEPS; Blossfeld & Roßbach, 2019), we examined $N = 511$ children with SEN-L from 93 classes in 56 special schools (“Förderschule”) and $N = 573$ children without SEN-L from 47 classes in 26 lower secondary schools (“Hauptschule”) attending fifth grade. The latter are low-achieving students without necessarily having learning difficulties because of below-average cognitive abilities. Average- or high-achieving students were not considered because the administered tests would have been substantially too easy for them and, thus, would not allow estimating reliable proficiency scores. Testing occurred in small groups at the respective schools by trained test administrators from a professional survey institute. The 1,084 children (48% girls) had a mean age of 11.31 years ($SD = 0.62$). Most of them (92%) were born in Germany. The compositions of the two samples were highly comparable in gender, age, and migration background (see Gnams & Nusser, 2022).

Instruments

Because the testing time was limited, only four subscales of the German version of the *Cognitive Abilities Test* (Heller & Perleth, 2000) could be administered. To cover figural as well as verbal item material, we selected two different domains measuring reasoning and verbal comprehension. All measures were presented as paper-based power tests (with generous time limits) and employed a number correct scoring scheme. For each subscale in the two domains, the maximum testing times were 9 and 7 min, respectively. Because these tests were developed for general student populations, we administered test versions that were designed for younger age groups (i.e., fourth grade). The reasoning was measured with the *figure classifications* and *figural analogies* subscales with 25 items each. The formerly presented items with three or four figures could be classified according to a distinct characteristic (e.g., form, shading, position). The children had to identify one out of five

figures that matched the classification of the target stimuli. The latter included items with pairs of figures that were logically related. For a target figure, the children had to identify one out of five figures that followed the same logical rule (i.e., analogy). Verbal comprehension was measured with 20 items of the *word analogies* subscale, each presenting a pair of words that were logically related. For a target word, the children had to identify one out of five words that followed the same logical association (i.e., analogy). Moreover, the *receptive vocabulary* subscale with 25 items required test takers to identify one out of five words corresponding to a written target word (e.g., synonyms). However, for the present analyses, we excluded Item 20 because preliminary analyses showed negative item discriminations in both groups.

Statistical Analyses

In line with the test authors, we fitted a unidimensional Rasch (1960) model with marginal maximum likelihood estimation to the responses of each test in the two samples that scored missing responses as incorrect. Item fit was evaluated using the weighted mean square (WMNSQ) statistic for which values of $WMNSQ < 1.15$ indicate close fit, $1.15 \leq WMNSQ < 1.20$ small misfit, and $WMNSQ \geq 1.20$ considerable misfit (e.g., Pohl & Carstensen, 2013). Residual diagnostics using Yen’s (1984) adjusted Q_3 statistic interpreted absolute values below .20 as an indication of essential unidimensionality. Significant model violations were detected using the test statistic proposed by Chalmers and Ng (2017).

Biases in single items and overall test scores in the form of differential item functioning (DIF) and differential test functioning (DTF) were examined by calculating the differences in the item and test score functions between children with and without SEN-L for each test. Following Chalmers (2018), these differences were given by the cDIF and cDTF statistics in the raw score metric. Thus, cDIF ranges between -1 and 1 , while the range of cDTF depends on the test length (e.g., for a test with 25 items cDTF falls between -25 and 25). Positive values indicate that children with SEN-L receive, on average, lower item or test scores than those without, despite holding the latent proficiency in both groups comparable. In contrast, negative values would indicate lower scores for children without SEN-L. To provide a comparable metric for the different tests, we also report the percentage bias cDTF% that gives the relative difference in test scores.

Results and Discussion

Although some children did not provide answers to all items, large missing rates were rare (see Figure 1).

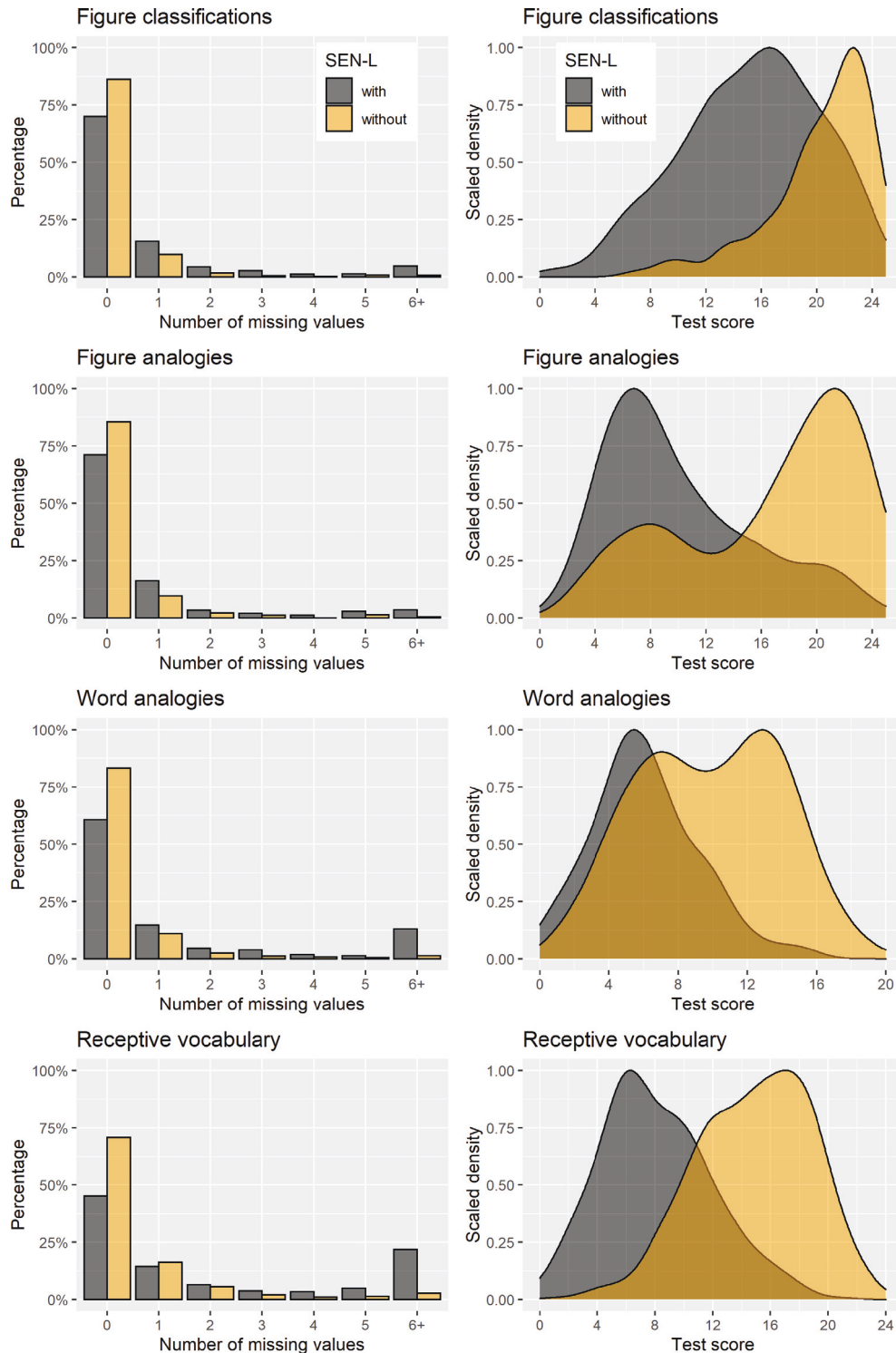


Figure 1. Distributions of missing responses and test scores for children with and without SEN-L.

However, children with SEN-L exhibited slightly more missing responses as compared to children without SEN-L. These differences were smaller for the two reasoning tests, Cohen's d s of 0.35, 95% CI [0.22, 0.48] and 0.32, 95% CI [0.19, 0.45], as compared to the two verbal

comprehension tests, Cohen's d s of 0.481, 95% CI [0.32, 0.65] and 0.67, 95% CI [0.54, 0.80] (see Gnamb & Nusser, 2022, for details). Moreover, the percentage of missing values for each item correlated between $r = .41$ and $r = .86$ with the item position. Descriptive comparisons

Table 1. Rasch model fit and differential response functioning for children with and without SEN-L

	Figure classifications	Figure analogies	Word analogies	Receptive vocabulary
Number of items	25	25	20	24 ^f
Model parameters for children with SEN-L				
Latent proficiency: <i>M</i> (<i>SD</i>) ^a	0.00 (1.10)	0.00 (1.07)	0.00 (0.69)	0.00 (0.78)
Item difficulties: <i>M</i> (<i>SD</i>)	-0.65 (1.16)	0.50 (0.84)	1.00 (0.79)	0.86 (0.96)
Model parameters for children without SEN-L				
Latent proficiency: <i>M</i> (<i>SD</i>) ^a	0.00 (1.14)	0.00 (1.56)	0.00 (1.04)	0.00 (0.79)
Item difficulties: <i>M</i> (<i>SD</i>)	-2.15 (1.31)	-1.06 (1.01)	0.06 (1.12)	-0.69 (1.15)
Item fit for children with SEN-L				
Number of large (small) WMNSQ ^b	0 (0)	0 (1)	0 (0)	0 (1)
Number of large (significant) <i>Q</i> ₃ ^c	0 (2)	0 (2)	0 (0)	0 (2)
Marginal reliability	.80	.81	.76	.79
Item fit for children without SEN-L				
Number of large (small) WMNSQ ^b	0 (0)	4 (0)	0 (0)	0 (0)
Number of large (significant) <i>Q</i> ₃ ^c	0 (2)	0 (13)	0 (8)	0 (1)
Marginal reliability	.70	.79	.78	.80
Differential item functioning ^d				
<i>Mdn</i> (cDIF)	0.01	0.01	0.02	-0.06
First/third quartiles of cDIF	-0.04/0.04	-0.05/0.04	-0.08/0.05	-0.07/0.06
Max (cDIF)	0.08	0.13	0.17	0.18
Differential test functioning ^e				
cDTF	0.02	-0.17	-0.24	-0.27
95% CI for cDTF	-0.42/0.45	-0.76/0.43	-0.68/0.17	-0.82/0.26
cDTF%	0.07%	-0.68%	-1.21%	-1.11%

Note. ^aMean proficiency was fixed to 0 for model identification. ^bNumber of items with weighted mean square error > 1.15 (small) or > 1.20 (large). ^cNumber of items with average adjusted Yen's *Q*₃ greater than 0.20 with the number of significant ($p < .05$) misfit (Chalmers & Ng, 2017) in parenthesis. ^dMedian, first and third quartiles, and maximum of signed differential item functioning statistics (cDIF) across items (Chalmers, 2018). ^eSigned differential test functioning statistic (cDTF) and percentage bias in test scores (cDTF%; Chalmers, 2018). ^fOne item was excluded because of a negative discrimination in both samples. Item-level results are available in the online repository.

showed slightly larger correlations for children with SEN-L as compared to those without, but only for the two verbal comprehension tests and not the two reasoning tests (see Gnambs & Nusser, 2022). This might suggest somewhat differential speededness between the two groups for the verbal tests.

For both groups, the fit of the items to the Rasch model can be considered satisfactory (see Table 1). Although the inference tests identified some misfitting items, the size of the model violations was not severe, as indicated by the WMNSQ and the *Q*₃ statistics. Because the item response models were estimated by constraining the mean of the ability distributions to zero, the mean item difficulties for each test inform about the targeting of the tests, that is, whether the average difficulty of the test matched the proficiency distribution of the sample. Despite administering out-of-level tests, three of the four tests were too difficult for children with SEN-L. The mean item difficulties fell about 0.5–1.0 standard deviations above the mean proficiency. Only the figure classification subscale was somewhat too easy for them. In contrast, for children without SEN-L, the tests either matched the sample's mean ability

or were too easy. However, all tests captured substantial individual differences between children, as indicated by the variances of the latent proficiency distributions that fell between 0.69 and 1.56. The marginal reliabilities of the tests were generally good for all tests and fell between .70 and .81.

Comparisons between children with and without SEN-L require comparable measurement structures of the administered tests in both groups. However, small biases in item scores (i.e., DIF) were observed in all administered tests, with some items exhibiting biases up to 18% (see Table 1). However, generally large DIF effects exceeding a bias of 5% were rare. The cumulated cDIF effects across all items of a test, as reflected in the respective cDTF statistics, showed no significant ($p > .05$) DTF for any test. Rather, the biases in test scores fell at about 1% at the most and, thus, indicated comparable measurements for children with and without SEN-L. Overall, these results show that the administered tests can be used to interpret mean-level differences between the two groups.

As expected, the density distributions in Figure 1 show that children with SEN-L exhibited substantially lower

proficiencies as compared to children without SEN-L. Moreover, the out-of-level tests seemed more appropriate for children with SEN-L because they were slightly too easy for children without SEN-L and resulted in slight ceiling effects.

To sum up, the results show the adequacy of out-of-level testing for children with SEN-L with four subtests of the German version of the *Cognitive Abilities Test*. Although these tests were originally developed for children without SEN-L, they also functioned adequately for children with SEN-L. The lack of substantial DTF also suggests that comparisons between the two groups seem feasible. However, two unresolved challenges for applied practice remain. First, it is not all clear how to properly choose the appropriate test level (i.e., age group) for children with SEN-L. In the present study, for three of the four administered scales, easier tests were (i.e., designed for third graders) might have been more appropriate for children with SEN-L. It is also conceivable that the appropriate level is test specific and depends, among others, on the measured construct (e.g., lower levels for verbal domains) and test material. A given test level might not necessarily be applicable for all students with SEN-L of the same age group but rather depend on further individual characteristics.

The second problem from an applied perspective is the lack of proper norms for children with SEN-L. When administering out-of-level tests, only age norms for children without SEN-L are typically available. But this does not allow comparing children with SEN-L to their peers and thus impedes comparisons between children with and without SEN-L. It would be beneficial if test developers provided test versions designed for different age groups linked and placed the different measurements on a common scale. Alternatively, using computer-adaptive testing formats that include items designed for different grade levels and cognitive demands might allow fair comparisons between children with a range of proficiencies. So far, these approaches are not yet commonly used in typical psychological assessments.

References

- Anderson, D., Lai, C.-F., Alonzo, J., & Tindal, G. (2011). Examining a grade-level math CBM designed for persistently low-performing students. *Educational Assessment, 16*(1), 15–34. <https://doi.org/10.1080/10627197.2011.551084>
- Blossfeld, H.-P. & Roßbach, H.-G. (Eds.). (2019). *Education as a lifelong process: The German National Educational Panel Study (NEPS)*. Edition ZfE (2nd ed.). Springer.
- Bosson, M. S., Hessels, M. G. P., Hessels-Schlatter, C., Berger, J.-L., Kipfer, N. M., & Büchel, F. P. (2010). Strategy acquisition by children with general learning difficulties through metacognitive training. *Australian Journal of Learning Difficulties, 15*(1), 13–34. <https://doi.org/10.1080/19404150903524523>
- Chalmers, R. P. (2018). Model-based measures for detecting and quantifying response bias. *Psychometrika, 83*(3), 696–732. <https://doi.org/10.1007/s11336-018-9626-9>
- Chalmers, R. P., & Ng, V. (2017). Plausible-value imputation statistics for detecting item misfit. *Applied Psychological Measurement, 41*(5), 372–387. <https://doi.org/10.1177/0146621617692079>
- Fuchs, L. S., Fuchs, D., & Capizzi, A. M. (2005). Identifying appropriate test accommodations for students with learning disabilities. *Focus on Exceptional Children, 37*(6), 1–8. <https://doi.org/10.17161/fec.v37i6.6812>
- Gnams, T., & Nusser, L. (2019). The longitudinal measurement of reasoning abilities in students with special educational needs. *Frontiers in Psychology, 10*, Article 232. <https://doi.org/10.3389/fpsyg.2019.00232>
- Gnams, T., & Nusser, L. (2022). *Out-of-level testing of children with SEN-L* [Computer code]. <https://osf.io/mwrpv/>
- Grünke, M., & Grosche, M. (2014). Lernbehinderung [Learning disability]. In G. W. Lauth & M. Grünke (Eds.), *Interventionen bei Lernstörungen* [Intervention in learning disabilities] (pp. 76–89). Hogrefe.
- Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision* [Cognitive Abilities Test for 4th to 12th grade, revision]. Beltz.
- Lloyd, J. W., Keller, C., & Hung, L. (2007). International understanding of learning disabilities. *Learning Disabilities Research and Practice, 22*(3), 159–160. <https://doi.org/10.1111/j.1540-5826.2007.00240.x>
- MacLean, H., McKenzie, K., Kidd, G., Murray, A. L., & Schwannauer, M. (2011). Measurement invariance in the assessment of people with an intellectual disability. *Research in Developmental Disabilities, 32*, 1081–1085. <https://doi.org/10.1016/j.ridd.2011.01.022>
- Minnema, J. E., Thurlow, M. L., Bielinski, J., & Scott, J. K. (2001). Past and current research on out-of-level testing of students with disabilities. *Assessment for Effective Intervention, 26*(2), 49–55. <https://doi.org/10.1177/073724770102600208>
- Nusser, L., & Weinert, S. (2017). Appropriate test-taking instructions for students with special educational needs. *Journal of Cognitive Education and Psychology, 16*(3), 227–240. <https://doi.org/10.1891/1945-8959.16.3.227>
- Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study. *Journal for Educational Research Online, 5*, 189–216. <https://doi.org/10.25656/01:8430>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145. <https://doi.org/10.1177/014662168400800201>

History

Received March 9, 2022

Revision received July 22, 2022

Accepted August 18, 2022

Published online November 2, 2022

EJPA Section / Category Educational Psychology

Conflict of Interest

The authors are employed at the Leibniz Institute of Educational Trajectories (LifBi, Germany) that conducts the NEPS. However, the institute had no involvement in the analyses of the data or the writing of the manuscript. The authors received no financial benefits for the publication of the manuscript.

Open Science

This paper uses data from the National Educational Panel Study (NEPS; see Blossfeld & Roßbach, 2019). The NEPS is carried out by the Leibniz Institute for Educational Trajectories (LifBi, Germany) in cooperation with a nationwide network.

Open Data: Due to legal restrictions, the information needed to reproduce all of the reported results are not openly accessible.

Open Materials: We confirm that there is sufficient information for an independent researcher to reproduce the reported methodology (Gnambs & Nusser, 2022).

Preregistration of Studies and Analysis Plans: This study was not preregistered.

Funding

Open access publication enabled by the University of Bamberg.

ORCID

Timo Gnambs

 <https://orcid.org/0000-0002-6984-1276>

Timo Gnambs

Leibniz Institute for Educational Trajectories

Wilhelmsplatz 3

96047 Bamberg

Germany

timo.gnambs@lifbi.de