



# Comparing Perceptual Speed Between Educational Contexts

## The Case of Students With Special Educational Needs

Timo Gnams<sup>1</sup>, Anna Scharl<sup>1</sup>, and Theresa Rohm<sup>1</sup>

Leibniz Institute for Educational Trajectories, Bamberg, Germany

**Abstract.** Perceptual speed is a basic component of cognitive functioning that allows people to efficiently process novel visual stimuli and quickly react to them. In educational studies, tests measuring perceptual speed are frequently developed using students from regular schools without considering students with special educational needs. Therefore, it is unclear whether these instruments allow valid comparisons between different school tracks. The present study on  $N = 3,312$  students from the National Educational Panel Study evaluated differential item functioning (DIF) of a short test of perceptual speed between four school tracks in Germany (special, basic, intermediate, and upper secondary schools). Bayesian Rasch Poisson counts modeling identified negligible DIF that did not systematically disadvantage specific students. Moreover, the test reliabilities were comparable between school tracks. These results highlight that perceptual speed can be comparably measured in special schools, thus enabling educational researchers to study schooling effects in the German educational system.

**Keywords:** perceptual speed, differential item functioning, large-scale assessment, Poisson counts models, Bayesian analyses

Educational research frequently involves comparisons between different school tracks, for example, to evaluate the effect of different curricula or instructional approaches on academic achievement. To properly address these research questions, the administered measures must be comparable across contexts. Otherwise, comparisons are unfair and might lead to biased conclusions. In practice, systematic differences in the measurement properties of cognitive tests might be expected if tests are administered in contexts they have not been developed or validated for. Many commercially available tests and even custom-designed tests administered in educational large-scale assessments are developed using students attending regular schools. However, if these tests are also administered to students with cognitive impairments (e.g., with special educational needs [SENs]), the measured constructs might differ to some degree, for example, because these students process instructions or item contents differently than regular students (Nusser & Weinert, 2017; Pohl et al., 2016; Südkamp et al., 2015). Before comparing a measured cognitive ability between different educational contexts, measurement invariance must be demonstrated. Therefore, the present study evaluates differential item functioning (DIF) in a test for figural perceptual speed. Of particular interest are students with SENs in the area of learning who attend various special schools in Germany (cf. Heydrich et al., 2013). Using a latent variable approach in a Bayesian framework, we examined whether the test

allows for fair comparisons between different school tracks.

## Theoretical Background

### Perceptual Speed as a Facet of Processing Speed

Processing speed is a component of cognitive functioning and represents the ability to quickly identify, discriminate, and decide about visual, auditory, or kinesthetic sensory information of different types of complexity (Holdnack, 2019). Measures of processing speed indicate how efficiently an individual can perform basic tasks in early stages of information processing. Slow processing speed may make a cognitive task (e.g., solving a math problem) more difficult, whereas higher speed can support thinking and learning. In the three-stratum model of cognitive abilities (Carroll, 1993), processing speed represents one of the eight broad abilities of which perceptual speed is one narrow facet. Perceptual speed indicates the automaticity and efficiency of processing novel visual information and the speed of decision-making. It is typically assessed with speeded tests that require the quick identification of specific targets from a set of stimuli. In these tests, perceptual speed is quantified either as the time until all targets are identified

or as the number of correctly identified targets per time. Perceptual speed is routinely assessed in various contexts because of its ability to predict various real-life outcomes. For example, in occupational and educational settings, it was associated with job (Mount et al., 2008) and school performance (Rindermann & Neubauer, 2004). Moreover, meta-analytic evidence highlighted the importance of processing speed for learning because children and adults with mathematical difficulties or reading disorders typically show substantially impaired speed performance compared to healthy comparison groups (Kudo et al., 2015; Peng et al., 2018). Thus, individual differences in perceptual speed can have a profound impact on learning outcomes and academic success.

## The Educational System in Germany

Germany has a tiered system of educational tracking that separates children at the age of about 10 years by ability into different school tracks. Low-achieving students attend *Hauptschule* (basic secondary school) and receive simplified educational training up to the ninth school grade, while students in *Realschule* (intermediate secondary school) receive extended education combined with more practical elements that may lead to an apprenticeship after tenth grade. High-achieving students attend *Gymnasium* (upper secondary school) and receive more advanced instructions in the same academic subjects and qualify for university entrance after the twelfth grade. Exceptions are students with cognitive difficulties and, therefore, SENs, for example, in the area of learning (SEN-L). These have problems in comprehending complex and abstract information which frequently leads to performance difficulties in regular schools (e.g., Nusser & Weinert, 2017). Moreover, SEN-L as compared to regular students shows frequently impaired cognitive performance in various domains such as reasoning abilities (Gnambs & Nusser, 2019), different components of working memory (Pickering & Gathercole, 2004), or reading competencies (Pohl et al., 2016). Therefore, students with SEN-L typically attend *Förderschule* (special school) that provides training and support targeted at the difficulties of these students.

## Comparison of Cognitive Abilities Between Educational Contexts

Evaluating the impact of different educational contexts on student outcomes requires valid and fair assessments of cognitive abilities in the examined contexts. Otherwise, test scores may not be comparable. However, cognitive

tests are typically developed using samples from regular schools that rarely include SEN-L students. But those students might have more difficulties in appropriately understanding standard test instructions and testing procedures (Nusser & Weinert, 2017). Similarly, they might interpret item contents differently or adopt less effective task solution strategies. All of this can contribute to differential test functioning between school tracks and result in incomparable measurements.

The few studies that evaluated the measurement properties of cognitive tests among SEN-L students concluded that comparative analyses are difficult or even impossible because the administered tests seemed to measure different constructs in different educational contexts (Bolt & Ysseldyke, 2008; Pohl et al., 2016; Südkamp et al., 2015). For example, a test for mathematical competence showed substantial DIF between groups of students with different SENs (Bolt & Ysseldyke, 2008). Similarly, Südkamp et al. (2015) showed that a valid comparison of reading competencies between students from regular and special schools was impossible because of substantial rates of missing responses, low item discrimination, and an inferior test reliability among SEN-L students. In contrast, Gnambs and Nusser (2019) reported that a short instrument measuring reasoning abilities exhibited comparable measurement properties among students from special and regular schools. Thus, the matter of measurement invariance across educational contexts seems to be test-specific and needs to be explored for each test setting anew.

## Present Study

The comparison of cognitive abilities across educational contexts requires comparable measurements in the studied settings. Therefore, the present study examines DIF of a test measuring figural perceptual speed between students from four educational tracks in Germany. Of particular interest are students with SEN-L who attend special schools because these were not considered during the development of the instrument (Lang et al., 2014). If the learning difficulties of students in special schools affect how they process and respond to the test items, comparisons between school tracks might be distorted. Like other frequently employed tests of cognitive speed (cf. Schmitz & Wilhelm, 2019), the present study evaluated an economical instrument that can be administered in less than 2 minutes and, thus, is ideally suited for educational large-scale studies where assessment times are costly. The test was limited to figural item material and, thus, is intended as a quick screening instrument to study population effects rather than a broad measure of mental speed for precise individual assessments.

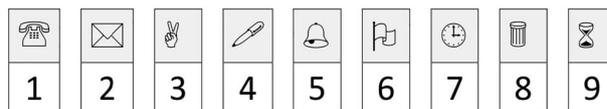
## Materials and Methods

### Sample and Procedure

The *National Educational Panel Study* (NEPS; Blossfeld et al., 2011) is a multicohort, large-scale assessment of student characteristics and educational outcomes in Germany. We draw on starting cohort 4 of the NEPS that included 11,580 students. To reduce confounding effects from systematic differences in the students' background characteristics, the students were matched across school tracks. While the matching worked well for sex, age, and migration background, it could not improve the distribution of the cultural capital indicator (for details, see Electronic Supplementary Material 1 [ESM 1]). Thus, the present study examined a matched sample of  $N = 3,312$  (45% girls) students from ninth grade attending 396 special ( $n = 901$ ), basic ( $n = 818$ ), intermediate ( $n = 789$ ), and upper secondary schools ( $n = 804$ ). Students attending specialized school types such as comprehensive schools (*Gesamtschule*) were not considered because of smaller sample sizes in these groups. Their  $M_{age}$  was 15.9 years ( $SD = 0.6$ ). All students were tested in small groups at their respective institutions by experienced supervisors who received a priori training to guarantee standardized assessment conditions.

### Measure

Figural perceptual speed was measured with three items from the *Bilder-Zeichen-Test* (Lang et al., 2014). For each



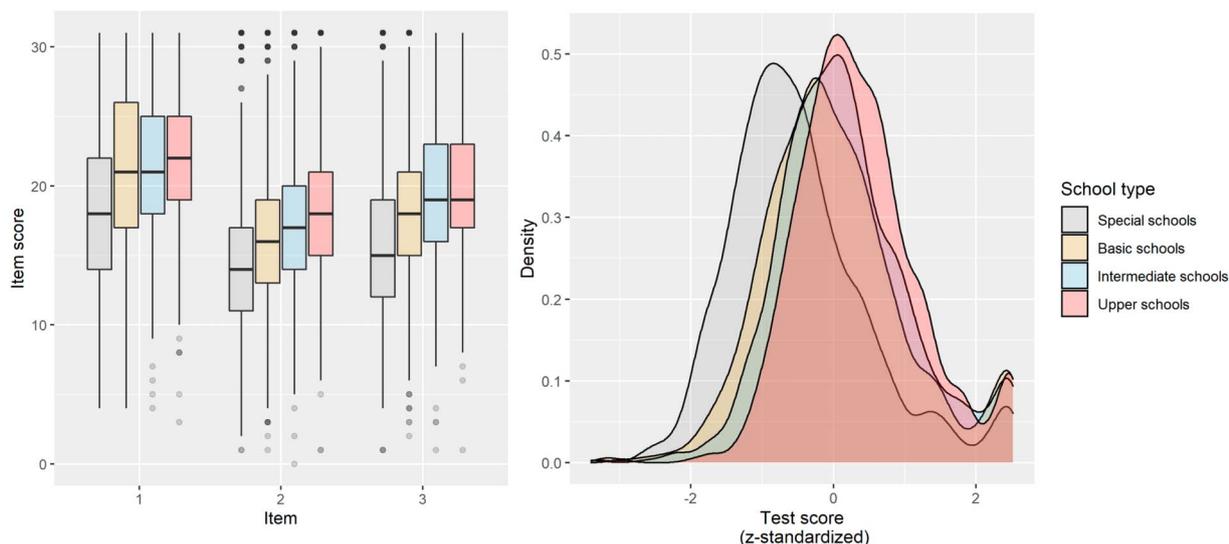
**Figure 1.** Target stimuli of an example item for the perceptual speed test (Lang et al., 2014, p. 9). © Leibniz Institute for Educational Trajectories (LifBi). Reproduced with permission.

item, a set of nine target stimuli including a figure and a corresponding number was presented on the top of the page (see Figure 1). Beneath the targets, 31 figures had to be matched to their target within 30 seconds by noting the number corresponding to the respective target stimulus. The number of correctly matched figures represented the item score (see left plot in Figure 2). The density plots of the  $z$ -standardized test scores in Figure 2 showed a distribution with two modes for each school track. Despite a time constraint of 30 seconds, the test exhibited ceiling effects for a non-negligible part of the sample. For methodological reasons (see below), we thus decided to reverse-code the items to indicate the number of errors (instead of correct responses).

### Statistical Analyses

#### Item Response Model

The total error scores of each item were modeled using the Rasch (1960) Poisson counts model parameterized as a generalized linear mixed-effects model (GLMM; Fox, 2010). In this approach, item difficulty parameters are represented by fixed effects, while person abilities are



**Figure 2.** Item and test score distributions by school track. Box plots of item scores are given on the left, while kernel density estimates of the  $z$ -standardized test score distributions are given on the right. Descriptive statistics for these distributions are reported in ESM 1.

given by a random effect. Because students were nested in different schools, we included an additional random school effect. To account for the second mode of the response distribution (see Figure 2), the model was extended by a zero-inflation process that accounts for the excess zeros of the error scores (see Lambert, 1992). As a robustness check, we compared several response distributions: (a) a Poisson distribution (without zero-inflation), (b) a zero-inflated Poisson counts distribution, (c) a zero-inflated Poisson-lognormal distribution, and (d) a zero-inflated Poisson-Gamma distribution. All but the Poisson-Gamma distribution were truncated at 31, the maximum number of errors in the administered items, because software constraints prevented us from truncating the Poisson-Gamma distribution. The dispersion parameter  $\varphi$  was calculated from (d) following Doebler and Holling (2016) to check if it substantially deviated from 1 (as implied by the Poisson counts model). Model comparisons were based on the Watanabe–Akaike information criterion and leave-one-out cross-validation (LOO; Vehtari et al., 2017) for which lower values indicate better fit. Overlapping confidence intervals for these indices were interpreted as comparable fit. The reliability of the scale along the proficiency distribution was estimated following Baghaei and Doebler (2019).

### Differential Item Functioning

In the GLMM framework, DIF is represented by significant cross-level interactions between the item parameters and some grouping variable (i.e., school track) after accounting for the main effects (Bürkner, 2020). In our application, both the probability for excess zeros and the rate parameter of the Poisson counts distribution can be regressed on covariates and, thus, allow for the examination of DIF. Therefore, we compared different models: (a) a model that included only main effects of school track and assumes no DIF, (b) a model with school track-specific DIF for the zero-inflation and item difficulty parameters, and (c) a model that additionally included school-level random item effects in both parts of the model. The latter allowed for distinguishing further differences stemming from the individual school's context (Hartig et al., 2020). The formal model specifications are given in ESM 1.

DIF effects were evaluated using 95% credible intervals (CrIs); effects that did not include 0 indicated DIF. The meaningfulness of the identified DIF was evaluated using a Cohen's *d*-like measure by standardizing the difference in item difficulties between two school tracks at their pooled population variances. Following the Educational Testing Service (Holland & Wainer, 1993), absolute values up to 0.25 (i.e., a quarter of a SD) were considered negligible, while values exceeding 0.50 were viewed as substantial DIF.

### Bayesian Model Estimation

All models were estimated using the R package *brms* (Bürkner, 2017) that provides an interface to Stan's (Carpenter et al., 2017) implementation of a Hamiltonian Monte Carlo algorithm. Fixed effects (e.g., the item difficulties and school track main effects) were modeled with weakly informative priors whose density mostly covered the commonly expected parameter space of the effects (see ESM 1 for details). For the DIF models, 8,000 posterior draws were obtained for each parameter. Convergence diagnostics for the focal models are reported in ESM 1. The posterior distributions of the parameters are summarized using the median and a 95% highest density interval to determine whether the parameter was different from zero.

### Open Practices

The anonymized data including information on the assessment procedure are available after registration at <https://doi.org/10.5157/NEPS:SC4:10.0.0>. Moreover, the R code including the analysis results is available at <https://osf.io/yfepc>.

## Results

The distributions of the *z*-standardized number correct scores for all school tracks substantially overlapped (see Figure 2). However, the mode for SEN-L students was markedly shifted to the left, while students in upper secondary schools had, on average, the highest test scores. First, we fitted four different item response models to the data without acknowledging differences between school tracks. Model comparisons (see Table 1) showed that acknowledging a zero-inflation process improved model fit compared to the ordinary Poisson counts model. Because differences between the selected response distributions were negligible, and little overdispersion was observed ( $\varphi = 1.02$ , 95% CrI [1.01, 1.03]), we proceeded with the zero-inflated Poisson counts distribution model. The item difficulty parameters (on the log-scale) were 2.33, 95% CrI [2.31, 2.36], 2.65, 95% CrI [2.62, 2.67], and 2.52, 95% CrI [2.49, 2.54] for the three items which corresponded to error scores (on a scale from 0 to 31) of 10.28, 14.12, and 12.37, respectively. Thus, the first item was slightly easier than the other two. For the zero-inflation process, the item parameters (on the logit-scale) were  $-4.13$ ,  $-6.19$ , and  $-5.42$ . Hence, the probability of observing a ceiling effect was only about 1% for item 1 and less than half this size for the remaining items. Given the negligible size of the zero-inflation parameters, we focus on DIF for the rate parameters of the Poisson process indicating the item difficulties.

**Table 1.** Model comparisons of estimated item response models

Response models	WAIC	SE <sub>WAIC</sub>	$\Delta_{WAIC}$	SE <sub><math>\Delta_{WAIC}</math></sub>	LOO	SE <sub>LOO</sub>	$\Delta_{LOO}$	SE <sub><math>\Delta_{LOO}</math></sub>
Models without DIF and different response distributions								
Poisson counts distribution without zero-inflation	60,370	293	4,777	251	60,912	301	5,007	255
Zero-inflated Poisson counts distribution	55,593	213	—	—	55,923	216	18	10
Zero-inflated Poisson-lognormal distribution	55,595	213	2	5	55,920	215	15	9
Zero-inflated Poisson-Gamma distribution	55,608	212	15	7	55,905	214	—	—
Zero-inflated Poisson counts distribution models with DIF								
Model 1: Model without DIF	55,582	214	859	60	55,925	217	702	61
Model 2: DIF model with school track-specific DIF	55,504	214	781	57	55,859	218	636	59
Model 3: DIF model with school-level random items effects	54,723	198	—	—	55,224	202	—	—

Note. DIF = differential item functioning, LOO = leave-one-out cross-validation, WAIC = Watanabe–Akaike information criterion,  $\Delta$  = difference to the lowest WAIC/LOO.

Next, we compared three models that acknowledged differences between school tracks (see Table 1). Model comparisons showed the best fit for the most complex model with fixed-effects DIF for the item parameters and, additionally, random group DIF across schools for each item. The estimated model parameters for the Poisson process are summarized in Table 2. The item difficulty parameters for the three perceptual speed items were similar to the previously reported results, with item 1 being the easiest and item 2 the most difficult. This pattern was rather robust and emerged for all four school tracks.

As expected, we observed mean differences (on the log scale) in the latent proficiencies between school tracks with respondents in regular schools exhibiting substantially larger perceptual speed compared to students in special schools,  $\beta = -.16$ , 95% CrI  $[-.22, -.10]$  for basic secondary schools,  $\beta = -.26$ , 95% CrI  $[-.32, -.19]$  for intermediate secondary schools, and  $\beta = -.29$ , 95% CrI  $[-.35, -.23]$  for upper secondary schools. These mean differences corresponded to standardized effect sizes  $\Delta$  of  $-0.62$ ,  $-0.99$ , and  $-1.16$ , respectively. Since we modeled error scores, the negative effects indicate fewer errors in, for example, upper secondary schools than in special schools and, thus, a higher proficiency. In contrast, the variances of the proficiency distributions did not differ markedly between school tracks with SDs of 0.23, 0.29, 0.27, and 0.24, respectively (see Table 2).

Moreover, the DIF model contained evidence for DIF for SEN-L students. Using item 3 as anchor item, item 1 was easier in basic schools than in special schools ( $\beta = -.07$ , 95% CrI  $[-.14, -.01]$ ), while item 2 was more difficult in basic schools ( $\beta = .05$ , 95% CrI  $[0.01, 0.09]$ ) and intermediate schools ( $\beta = .07$ , 95% CrI  $[0.03, -0.11]$ ). These DIF effects corresponded to about 0.93 to 1.17 percentage changes in mean scores and standardized ESs  $\Delta$  of  $-0.27$ , 0.19, and 0.28, respectively. Following rules of thumb for the interpretation of these ESs, two of them can

be considered as exhibiting moderate DIF. For the remaining school tracks, no substantial DIF was observed (see Table 2). Recently, it has been argued (Hartig et al., 2020) that, in addition to average DIF across fixed groups (i.e., school tracks), random group DIF should be studied to determine whether relevant differences in item difficulties exist between schools. The respective random school effects for item 2,  $\sigma = .03$ , 90% CrI  $[0.00, 0.06]$ , and item 3,  $\sigma = .05$ , 90% CrI  $[0.01, 0.08]$ , were rather small. Moreover, the posterior probabilities that these variances equaled 0 were 98% and 88%, thus giving weak evidence for random school DIF. In contrast, for item 1, the respective effect was substantially larger,  $\sigma = .17$ , 90% CrI  $[0.15, 0.19]$ , with a posterior probability of no variance of 0%, which suggests differences in the difficulty of item 1 between schools. These results also replicated in sensitivity analyses using unmatched samples across school types (see ESM 1).

Finally, we explored the reliability of the administered test in the four school tracks. The average reliability coefficients were .74 in special schools, .78 in basic secondary schools, .75 in intermediate secondary schools, and .73 in upper secondary schools. These model-based reliability estimates were slightly smaller than traditional omega reliabilities of .80, .80, .82, and .76, respectively. Moreover, the reliabilities along the proficiency scale are given in Figure 3. Although students with lower ability exhibited slightly lower reliabilities, the reliability estimates were reasonably high for a range of abilities and, typically, exceeded .70. Importantly, reliabilities did not differ substantially between school tracks.

## Discussion

Measurement is a cornerstone of educational and psychological research. If tests are not comparable across

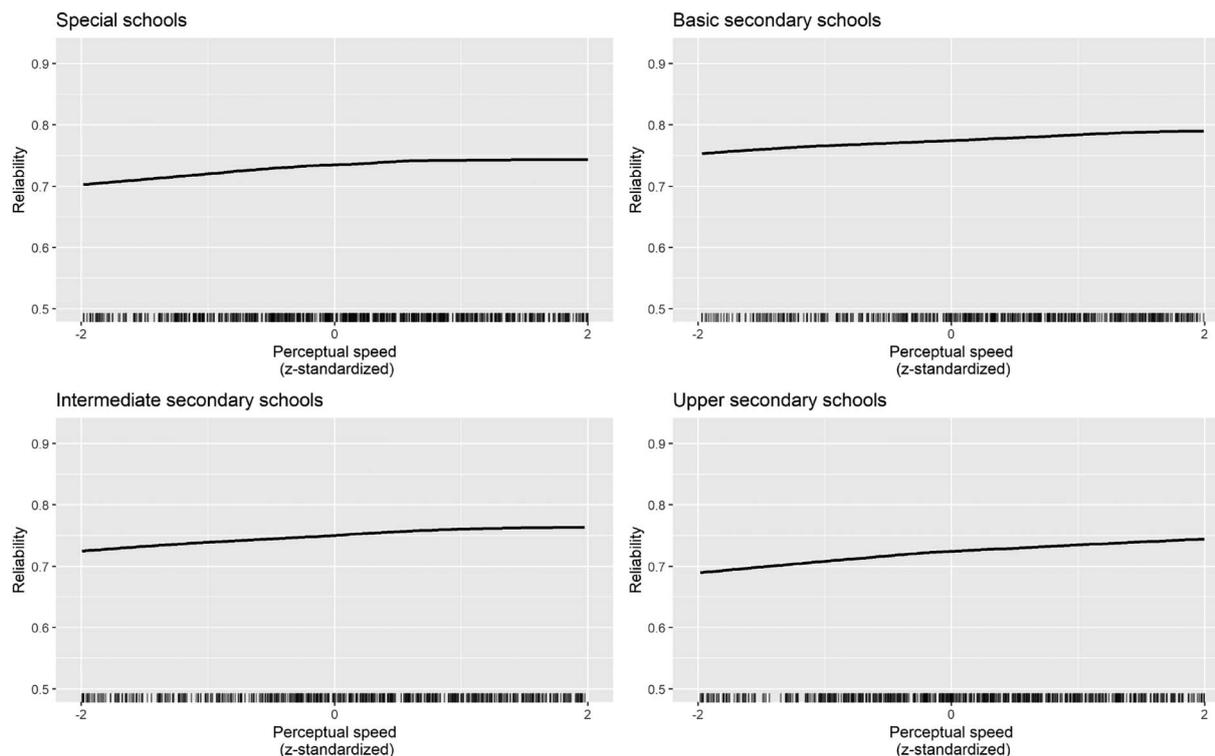
**Table 2.** Summary of model parameters for DIF effects

Parameters	<i>Mdn</i>	<i>MAD</i>	<i>LL CrI</i>	<i>UL CrI</i>	$e^{Mdn}$
Fixed effects					
Item 1	<b>2.53</b>	0.03	2.47	2.58	12.50
Item 2	<b>2.79</b>	0.02	2.74	2.83	16.22
Item 3	<b>2.69</b>	0.02	2.64	2.73	14.71
School track (ref. cat.: Special schools)					
BS	<b>-0.16</b>	0.03	-0.22	-0.10	0.85
IS	<b>-0.26</b>	0.03	-0.32	-0.19	0.77
US	<b>-0.29</b>	0.03	-0.35	-0.23	0.75
DIF effects (ref. cat.: Special schools, item 3)					
Item 1 × BS	<b>-0.07</b>	0.03	-0.14	-0.00	0.93
Item 2 × BS	<b>0.05</b>	0.02	0.01	0.09	1.05
Item 1 × IS	-0.01	0.03	-0.08	0.05	0.99
Item 2 × IS	<b>0.07</b>	0.02	0.03	0.11	1.07
Item 1 × US	-0.04	0.03	-0.10	0.03	0.96
Item 2 × US	0.03	0.02	-0.01	0.07	1.03
Random effects ( <i>SD</i> )					
Students					
Special schools	<b>0.23</b>	0.01	0.21	0.25	
BS	<b>0.29</b>	0.01	0.27	0.31	
IS	<b>0.27</b>	0.01	0.25	0.30	
US	<b>0.24</b>	0.01	0.22	0.26	
Schools					
Special schools	<b>0.18</b>	0.02	0.15	0.22	
BS	<b>0.13</b>	0.02	0.09	0.16	
IS	<b>0.16</b>	0.02	0.12	0.20	
US	<b>0.12</b>	0.02	0.09	0.15	
Items in schools					
Item 1	<b>0.17</b>	0.01	0.15	0.19	
Item 2	0.03	0.02	0.00	0.07	
Item 3	<b>0.05</b>	0.02	0.01	0.08	

Note. BS = basic schools, CrI = credibility interval, DIF = differential item functioning, IS = intermediate school, *LL CrI* = lower limit of the 95% CrI, *MAD* = median absolute deviation of the posterior distribution, *Mdn* = median of the posterior distribution, *UL CrI* = upper limit of the 95% CrI; US = upper schools.  $e^{Mdn}$  = exponentiated parameter estimate indicating the expected percentage change for a unit change in the predictor. Values in boldface indicate fixed effects for which zero is not contained in the CrI or the posterior probabilities of a variance of 0 exceeds 90%. Full results are given in ESM 1.

relevant groups, DIF can bias cross-group comparisons and lead to inappropriate conclusions. Particularly, in representative large-scale assessments, it is important to show that the administered measures can be used to compare individuals across different contexts. Therefore, the present study evaluated a short test measuring figural perceptual speed. We examined whether the test allows valid comparisons between different school tracks in the German educational system. These analyses revealed only modest DIF effects between special and regular schools. Importantly, the direction of DIF did not systematically disadvantage a specific school track:

While item 1 was more difficult for SEN-L students, item 2 was easier for them. Thus, it is unlikely that differences in the measurement properties of the test would systematically bias school track comparisons in substantial research. Interestingly, item-specific random school DIF was more pronounced (cf. Hartig et al., 2020). Particularly for the first item, notable differences in the estimated item difficulties were identified across schools. This might be a sign of problems in the standardization of the test procedure. Despite elaborated test protocols and extensive training of the test administrators, test instructions or organization of the test setting (e.g., the



**Figure 3.** Reliability by school track across the proficiency scale.

seating of students, promoting test engagement, and adherence to time limits) might have varied between schools to some degree and, thus, affected student performance at the beginning of the test. This underscores the importance of standardized assessment conditions in educational large-scale assessments for comparable cognitive measures across different educational contexts. Finally, the test exhibited acceptable levels of reliabilities across different levels on the proficiency scale. Importantly, no substantial differences in the measurement precision were observed for SEN-L students.

Despite the encouraging findings, the generalizability of our results might be limited by solely relying on figural item material. Given the higher prevalence of dyslexia and dyscalculia among SEN-L students (e.g., Van der Veen et al., 2010), more pronounced DIF effects might be observed for instruments with numeric or verbal item material. Furthermore, even though a three items scale might be considered short, each item of the figural speed test provided counts data and was, thus, substantially more informative than binary correct/incorrect responses in typical power tests. Moreover, for unidimensional cognitive scales, test shortening does not systematically impair criterion validities or mean-group comparisons (Heene et al., 2014). Finally, recent advancements in the modeling of counts data have suggested alternative modeling

strategies that involve more realistic assumptions for empirical data (e.g., Forthman et al., 2020). Although software constraints prevented us from exploring these modeling strategies, we have little reason to believe that this would have substantially affected our findings, as model comparisons did not suggest substantial overdispersion in our data.

In summary, the reported results demonstrate that the administered test of perceptual speed can be validly used for school track comparisons in educational large-scale assessments. These results fall in line with recent research (e.g., Gnams & Nusser, 2019), showing that some measures originally developed for students in regular schools can be comparably administered to SEN-L students. However, we want to emphasize that these results do not render comparable analyses for other cognitive measures obsolete. For other tests requiring higher reading abilities or higher-order thinking, or adopting more complex response formats, comparisons between special and regular schools might be more challenging (cf. Bolt & Ysseldyke, 2008; Pohl et al., 2016; Südkamp et al., 2015). Future research is also encouraged to extend comparable analyses to other school types such as comprehensive or Waldorf schools. This would strengthen comparative educational research for schools with substantially different curricula and pedagogical concepts.

## Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at <https://doi.org/10.1027/2698-1866/a000013>

**ESM 1.** The ESM includes information on (a) the matching of samples across school types, (b) descriptive statistics, (c) formal model specifications, (d) the Bayesian estimation and statistical software, (e) the parameter estimates, and (f) sensitivity analyses.

## References

- Bürkner, P. (2020). *Bayesian item response modeling in R with brms and Stan*. arXiv Preprint. <https://arxiv.org/abs/1905.09501>
- Bürkner, P. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Baghaei, P., & Doeblner, P. (2019). Introduction to the Rasch Poisson counts model: An R tutorial. *Psychological Reports*, 122(5), 1967–1994. <https://doi.org/10.1177/0033294118797577>
- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (2011). Education as a lifelong process – The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft*, 14, 19–34. <https://doi.org/10.1007/s11618-011-0179-2>
- 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(2), 121–138. <https://doi.org/10.1177/0734282907307703>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1). <https://doi.org/10.18637/jss.v076.i01>
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.
- Doeblner, A., & Holling, H. (2016). A processing speed test based on rule-based item generation: An analysis with the Rasch Poisson counts model. *Learning and Individual Differences*, 52, 121–128. <https://doi.org/10.1016/j.lindif.2015.01.013>
- Forthmann, B., Günhe, D., & Doeblner, P. (2020). Revisiting dispersion in count data item response theory models: The Conway–Maxwell–Poisson counts model. *British Journal of Mathematical and Statistical Psychology*, 73(S1), 32–50. <https://doi.org/10.1111/bmsp.12184>
- Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer.
- Gnams, T., & Nusser, L. (2019). The longitudinal measurement of reasoning abilities in students with special educational needs. *Frontiers in Psychology*, 10, 232. <https://doi.org/10.3389/fpsyg.2019.00232>
- Gnams, T., Scharl, A., & Rohm, T., (2021). Comparing perceptual speed between educational contexts [Data set]. <https://osf.io/yfepc/>
- Hartig, J., Köhler, C., & Naumann, A. (2020). Using a multilevel random item Rasch model to examine item difficulty variance between random groups. *Psychological Test and Assessment Modeling*, 62(1), 11–27.
- Heene, M., Bollmann, S., & Bühner, M. (2014). Much ado about nothing, or much to do about something? Effects of scale shortening on criterion validity and mean differences. *Journal of Individual Differences*, 35(4), 245–249. <https://doi.org/10.1027/1614-0001/a000146>
- 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(2), 217–240.
- Holdnack, J. A., Prifitera, A., Weiss, L. G., & Saklofske, D. H. (2019). WISC-V and the personalized assessment approach. In L. G., Weiss, D. H., Saklofske, J. A., Holdnack, & A., Prifitera (Eds.), *WISC-VE clinical use and interpretation* (pp. 447–488). Academic Press. <https://doi.org/10.1016/B978-0-12-815744-2.00013-6>
- Holland, P. W., & Wainer, H. E. (1993). *Differential item functioning*. Erlbaum.
- Kudo, M. F., Lussier, C. M., & Swanson, H. L. (2015). Reading disabilities in children: A selective meta-analysis of the cognitive literature. *Research in Developmental Disabilities*, 40, 51–62. <https://doi.org/10.1016/j.ridd.2015.01.002>
- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1), 1–14. <https://doi.org/10.2307/1269547>
- Lang, F. R., Kamin, S., Rohr, M., Stünkel, C., & Williger, B. (2014). *Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen des Nationalen Bildungspanels* [Assessment of fluid cognitive abilities across the life span in the National Educational Panel Study] (NEPS Working Paper No. 43). Leibniz-Institute for Educational Trajectories.
- Mount, M. K., Oh, I.-S., & Burns, M. (2008). Incremental validity of perceptual speed and accuracy over general mental ability. *Personnel Psychology*, 61(1), 113–139. <https://doi.org/10.1111/j.1744-6570.2008.00107.x>
- 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>
- Peng, P., Wang, C., & Namkung, J. (2018). Understanding the cognition related to mathematics difficulties: A meta-analysis on the cognitive deficit profiles and the bottleneck theory. *Review of Educational Research*, 88(3), 434–476. <https://doi.org/10.3102/0034654317753350>
- Pickering, S. J., & Gathercole, S. E. (2004). Distinctive working memory profiles in children with special educational needs. *Educational Psychology*, 24(3), 393–408. <https://doi.org/10.1080/0144341042000211715>
- Pohl, S., Südkamp, A., Hardt, K., Carstensen, C. H., & Weinert, S. (2016). Testing students with special educational needs in large-scale assessments – Psychometric properties of test scores and associations with test taking behavior. *Frontiers in Psychology*, 7, 154. <https://doi.org/10.3389/fpsyg.2016.00154>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.
- Rindermann, H., & Neubauer, A. (2004). Processing speed, intelligence, creativity, and school performance: Testing of causal hypotheses using structural equation models. *Intelligence*, 32(6), 573–589. <https://doi.org/10.1016/j.intell.2004.06.005>
- Südkamp, A., Pohl, S., & Weinert, S. (2015). Competence assessment of students with special educational needs—Identification of appropriate testing accommodations. *Frontline Learning Research*, 3(2), 1–26. <https://doi.org/10.14786/flr.v3i2.130>
- Schmitz, F., & Wilhelm, O. (2019). Mene mene tekel upharsin: Clerical speed and elementary cognitive speed are different by virtue of test mode only. *Journal of Intelligence*, 7(3), 16. <https://doi.org/10.3390/jintelligence7030016>
- Van der Veen, I., Smeets, E., & Derriks, M. (2010). Children with special educational needs in the Netherlands: Number, characteristics and school career. *Educational Research*, 52(1), 15–43. <https://doi.org/10.1080/00131881003588147>
- Vehtari, A., Gelman, A. & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>

**History**

Received February 19, 2021  
Revision received June 30, 2021  
Accepted July 1, 2021  
Published online August 27, 2021  
Section: Educational Psychology

**Acknowledgments**

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 9, doi:10.5157/NEPS:SC4:10.0.0. From 2008 to 2013, NEPS data was 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, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LifBi) at the University of Bamberg in cooperation with a nationwide network.

**Conflict of Interest**

The authors have no conflicts of interest to disclose.

**Open Data**

The anonymized data including information on the assessment procedure are available after registration at <https://doi.org/10.5157/NEPS:SC4:10.0.0>.

Moreover, the R code including the analysis results is available at <https://osf.io/yfecp> (Gnamb et al., 2021).

**Funding**

The publication of this article was funded by the Open Access Fund of the Leibniz Association.

**ORCID**

Timo Gnamb  
 <https://orcid.org/0000-0002-6984-1276>  
Anna Scharl  
 <https://orcid.org/0000-0003-0081-1893>  
Theresa Rohm  
 <https://orcid.org/0000-0001-9203-327X>

**Timo Gnamb**

Leibniz Institute for Educational Trajectories  
Wilhelmsplatz 3  
96047 Bamberg  
Germany  
[timo.gnamb@lifbi.de](mailto:timo.gnamb@lifbi.de)