Equivalence of Computer- and Paper-based Administrations of the SCHNAPP Spelling Test in German for Six-Year-Old Children

Martin Schöfl[1,2], Gabriele Steinmair[1], Sabine Zepnik[1], and Timo Gnambs[3]

[1] Department of Educational Sciences, University of Education Upper Austria, Austria

[2] Research Institute for Developmental Medicine, Johannes Kepler University Linz, Austria

[3] Leibniz Institute for Educational Trajectories, Germany

Accepted for publication in the *European Journal of Psychology*:

https://doi.org/10.1027/1015-5759/a000793

Abstract

The *SCHNAPP Spelling Test* is a novel screening instrument to identify at-risk children with poor spelling abilities in German at the beginning of primary school. Although originally developed as a computerized test to be administered on tablets, in school settings paper-pencil methods are often still preferred. Therefore, the present study on $N = 390$ children from first grades in Austrian primary schools examined the equivalence of computer and paper-pencil versions of the test. After demonstrating unidimensional measurement models in both assessment conditions, analyses of differential response functioning on the item and test level found no substantial testing mode effects. These results indicate that the *SCHNAPP Spelling Test* can be comparably used as a computer- or paper-based instrument in school assessments.

Keywords: mode effect, spelling test, writing, paper-pencil, digital assessment

Equivalence of Computer- and Paper-based Administrations of the SCHNAPP Spelling Test in German for Six-Year-Old Children

Spelling abilities are an important foundation for educational and occupational success (e.g., Pan et al., 2021). That is why already in the first grades of primary school, spelling competence emerges as an influential predictor of later-life abilities (e.g., Caravolas et al., 2001; Landerl & Wimmer, 2008; Mesquita et al., 2022). To prevent long-lasting disadvantages resulting from poor spelling skills, it is important to identify children with deficits in spelling abilities early in their school careers. This requires appropriate measurement instruments with sound psychometric properties. For that reason, the *SCHNAPP Spelling Test*[1] (see Schöfl et al., 2023) was recently developed to serve as an economic screening instrument of spelling abilities in German at the beginning of primary school. The test is typically administered individually on tablet to allow for highly standardized instructions and item presentations. However, for routine screening procedures at school, the use of a digital spelling tests is not always feasible because respective devices are not readily available in sufficient quantity. Therefore, the present study introduces a paper-based version of the spelling test that can also be administered in large group settings. We present systematic analyses of differential response functioning (see Chalmers, 2018) to evaluate the measurement equivalence between the new paper-based and the standard digital test version. These analyses demonstrate that the computer- and paper-based versions of the *SCHNAPP Spelling Test* allow for comparable measurements of early spelling abilities without introducing notable mode effects.

[1] The acronym for the test was derived from the German translation of the research project „Literacy Acquisition at the Interface with Primary Education“ that funded the test development.

Spelling Competencies as a Facet of General Literacy

Spelling ability in a given language not only refers to a person's ability to form words with the correct letters in their proper sequence, but also shows the level of insight into the writing system of a particular language. Spelling is considered an important building block of general literacy later in life. This influence becomes apparent, for example, through the development of orthographic awareness resulting in knowledge of typical orthographic patterns of an alphabetic writing system such as English and German (Cheema et al., 2023; Domahs et al., 2016). The importance of spelling abilities is also emphasized by developmental studies that highlight intraindividual connections between deficits in spelling and reading difficulties in different writing systems (De Sousa Lopes & Carvalho Bedulho, 2022; Cheema et al., 2023; Pan et al., 2021) or between spelling competence and handwriting speed which is considered an important determinant of literacy development (e.g., for Spanish orthography see Afonso et al., 2020). Thus, correct spelling is an essential requirement for effective reading. If the written form is not spelled correctly, the reader might not be able to extract its meaning at all or only after considerable time (Pan et al., 2021). Research dealing particularly with the German writing system points to the relevance of the written form for reading comprehension. Thus, the use of structural units such as syllables, the trochaic foot, and morphemes are essential for recoding words, while on the next level capital letters, spaces between words, and punctuation are required for proper recoding syntactical units to establish basic reading skills (e.g., Evertz & Primus, 2013; Neef, 2002). These structural units can also be useful for spelling skills (e.g., Domahs et al., 2001, for syllabic principles).

The SCHNAPP Spelling Test in German

Increasing evidence suggests that not only the strategy "listen carefully to the sounds of a word" in the sense of a phonographic realization, but also the simple learning of spelling rules to the effect that isolated cases are committed to memory is often not optimal for an efficient acquisition of spelling skills (see Pan et al., 2021). Rather, current linguistic research

favours an approach that takes structural effects of the writing system, such as syllabic and morphological principles, into account (see Fuhrhop & Peters, 2013, for the German writing system). These structural approaches better support children acquiring spelling skills by discovering patterns within the German writing system rather than learning example cases by heart (Bredel, 2015). This rather novel perspective served as basis for the recently developed *SCHNAPP Spelling Test* (Schöfl et al., 2023) that was conceived as an economic screening instrument to identify at-risk children in primary school with poor early spelling abilities in German. The test includes 22 words that were selected based on their distribution in the typical vocabulary corpus of young children (childlex; Schroeder et al., 2015). Words from the entire range were included to avoid biases due to more or less frequent words. These words were chosen according to a theoretical hierarchy of vocabulary in the German writing system based on the trochaic foot, a rhythmic sequence of a strong and weak syllable (Evertz & Primus 2013; Primus, 2010). Thus, the selected words correspond to *a priori* hypothesized difficulty levels that target a broad range of proficiencies. Previous research not only attested to the instrument's unidimensionality and measurement precision but also confirmed the assumed difficulty hierarchy of the items (see Schöfl et al., 2023).

The test is typically administered digitally on tablets with a digital pen that allow highly standardized item presentations and self-paced test progress by presenting the target words via headphones. Thus, the test follows a general trend in psychological and educational assessment towards an increasing integration of digital media in their practice (e.g., Wright, 2020; Zinn et al., 2021). Nevertheless, at school, the use of digital spelling tests is not always feasible, for example, because respective devices are not available, particularly when testing larger groups of children. Therefore, it would be convenient to have alternative paper-based assessment formats. However, to use the *SCHNAPP Spelling Test* interchangeably as a paper-based test (PBT) or computer-based test (CBT), depending on the current situational conditions, the testing mode must not affect the measured competencies, that is, measurement

invariance must hold. Otherwise, the test media would involuntarily influence the measurements of children's spelling abilities and, potentially, distort raw score comparisons.

Comparability of Testing Modes

Whether different assessment modes such as CBT or PBT systematically distort psychological measurements is still controversially discussed (e.g., Clinton, 2019; Gnambs & Lenhard, 2023; Wright, 2020; Zinn et al., 2021). Whereas early research, for example, on mode effects for tests of reading competence suggested that respondents achieved lower scores on computer as compared to paper (e.g., Clinton, 2019), more recent studies found approximate measurement invariance between testing modes (e.g., Gnambs & Lenhard, 2023). Potential reasons for these contradictory findings are that mode effects (if they exist) are test-specific or depend on respondent characteristics such a computer familiarity (see Lynch, 2022). Therefore, the question of comparable measurements needs to be addressed for each instrument and target group.

Specifically tests of orthographic skills need to consider the process of writing in addition to the media used to present the items. Until now, most studies explored differences between keyboard writing versus handwriting (e.g., Horkay et al., 2006; White et al., 2015). For example, Horkay and colleagues (2006) observed negligible mode effects on writing performance among 13- to 14-year-olds; albeit computer test performance increased with higher computer familiarity. In contrast, another study on fourth grade students found that low-performing students achieved lower scores on a computerized writing test using keyboards as compared to a traditional paper-based test using handwriting (White et al., 2015). For high-performing students a reverse effect was observed. A reason for the reported mode effects might be that the motoric skills required for handwriting are not comparable to typing on a keyboard. But recent technological advances allow creating more natural writing situations by using tablets with digital pens. This might result in writing processes that are highly comparable to traditional handwriting on paper. Although preliminary findings

indicate that young children seem to prefer the use of digital as compared to traditional pens (Mombach et al., 2020), so far, little is known whether handwriting on paper or with a digital pen affects writing performance.

Present Study

The *SCHNAPP Spelling Test* (Schöfl et al., 2023) is available as a digital version to be administered on tablets with digital pens and a traditional paper-based test that might be more suitable for assessment situations including large groups at school. Because previous research indicated that different testing modes might distort proficiency estimates and complicate cross-mode comparisons (e.g., Clinton, 2019; White et al., 2015), the present study evaluated the equivalence of the PBT and CBT in a sample of primary school children. Importantly, we did not expect pronounced mode effects but rather assumed that both versions can be used interchangeably because both test versions were constructed to be highly comparable by using tablets with digital pens that should closely resemble handwriting on paper.

Materials and Methods

Sample and Procedure

At the end of the school year 2020/21, we selected a convenience sample of 390 children (51% girls) attending 25 different first classes from 10 primary schools in Northern Austria. The project team asked a couple of rural schools and schools from the city, all of them participated. In each school, all first graders for which parental consent was available were eligible to participate. Their mean age was about 6 years. About 76% of them spoke only German at home, whereas the rest had a bilingual background and additionally used various languages such as English, Bosnian, or Russian. The children’s socioeconomic background was rather high, as indicated by the highest parental education. About 49% of the children had a parent with a university degree, whereas only 24% of the parents had lower secondary education or vocational training. As compared to official statistics (Statistik Austria, 2022), the present sample can be considered roughly representative in terms of sex (51% versus 48%

girls), but included slightly less children with non-German primary language at home (24% versus 31%).

The school classes were originally randomly assigned to the different assessment conditions to have equal subsamples. But, because of the corona pandemic some assessments, primarily in the CBT condition, could not be conducted. As a result, in 14 classes including 242 children, the PBT version of the spelling test was administered, whereas, in another 11 classes with 148 children, the CBT was presented[2]. The two groups did not differ significantly ($p > .05$) regarding their age, first language, or parental education (see Table 1).

Instrument and Procedure

All children received the *SCHNAPP Spelling Test* (Schöfl et al., 2023) with 22 items that were individually presented on a single page. The test mode was assigned by school class; consequently, all children within a class received the PBT or CBT. In the CBT condition, each child received a tablet (iPad, 17th generation, 10.2 inches), headphones, and a digital pen. The test procedure was explained within a motivating frame story. To practice using the digital pen and writing on a tablet, two tasks ("Write your own name!" and an unscored sample item) were presented beforehand. Children could repeat instructions and practice as often as desired. In the PBT condition, each child was given a test booklet with cloze texts, while the test administrator (i.e., trained students or members of the research group) dictated the target words for the entire class. If a child asked for a repetition of the target word, the item was repeated for the entire class. The same instruction and practice items as in the CBT condition were used.

[2] One child from a class assigned to the PBT condition was individually tested at a later time because it was ill at the original test date. This child was mistakenly administered the CBT version of the test and, thus, is included in the CBT sample.

Statistical Analyses

Following Schöfl and colleagues (2023), the spelling test was scaled using a one-parametric logistic item response model (Rasch, 1960). To demonstrate the adequacy of the measurement model, we compared the fit of the Rasch (1960) model that only acknowledges different item difficulties and the two-parametric item response model (Birnbaum, 1986) that also considers different discrimination parameters for each item. For model comparisons we relied on the sample-size adjusted Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978) which indicate a better fit at lower values. Moreover, we evaluated the fit of each item using the weighted mean squared error (WMNSQ; Wright & Masters, 1982) for which values smaller than 1.15 are considered satisfactory for operational use (Pohl & Carstensen, 2013). The inference test S-χ^2 (Orlando & Thiessen, 2000) was used to examine differences between the expected and observed item response curves.

Mode effects between the two assessment conditions were studied by examining differential response functioning for each item and the entire test. A test exhibits differential item or test functioning (DIF, DTF) when the expected item or test scores differ between the two groups despite comparable latent proficiencies (Penfield & Camilli, 2006). To place the different test versions on a common scale, we fitted a multi-group item response model to the data that constrained the item difficulties for three measurement invariant anchor items across groups. Moreover, the latent mean of the PBT group was fixed to 0 and, thus, acted as a reference group. In contrast, the remaining item parameters, the latent population variances in both groups, and the latent mean of the CBT group were freely estimated. DIF is often viewed as problematic when the difference in item difficulties exceeds .43 or .63 which are often considered medium or large differences, respectively (Penfield & Camilli, 2006). As a more precise indicator of DIF, we also report the bias-corrected root mean squared deviation (RMSD; Köhler et al., 2020) that summarizes the differences between the group-wise item

response functions and the common item response function based on all respondents. Negligible DIF is reflected by values smaller than .08 (Robitzsch & Lüdtke, 2020). Moreover, the weighted root mean squared deviation (WRMSD), which combines the two group-wise RMSD statistics into a single weighted average, indicates noteworthy DIF for values exceeding 0.15 (Li, 2012). Subsequently, biases in test scores were studied using the cDTF statistic by Chalmers (2018) that gives the differences in the test score functions between the two groups and, thus, can fall between -22 and 22 in our case. Negative values indicate that the PBT received, on average, lower test scores than the CBT, despite comparable latent proficiency in both groups. The hypothesis of equivalent test scores in the two mode conditions was evaluated using an adapted inference test for cDTF (see Chalmers, 2018) that followed the TOST (two one-sided tests) procedure to test for equivalence (see Lakens et al., 2018). To this end, we considered differences in expected test scores in the range of ±0.88 points (which corresponds to about 0.2 units on the *z*-standardized test score scale) as equivalence indicating no mode effects.

Software

The analyses were conducted in *R* (Version, 4.2.3; R Core Team, 2023). The item response models including analyses of differential response functioning were estimated with the *mirt* package (Version 1.38.1; Chalmers, 2012). The bias-corrected RMSD statistics were derived using TAM (Version 4.1-4; Robitzsch et al., 2022).

Results

Measurement Models for Test Versions

The appropriate item response model was identified by comparing the fit of the Rasch model and the two-parametric logistic model. As summarized in Table 2, for the paper-based spelling test, both the AIC and the BIC showed a better fit for the simpler Rasch (1960) as compared to the more complex two-parametric item response model. Similarly, the BIC also preferred the Rasch (1960) model for the computer-based test. However, the AIC was slightly

better for the more complex model. Because the Rasch (1960) model underlies the theoretical rationale that guided the test development and conforms to the scoring rules as outlined by the test authors (see Schöfl et al., 2022), we chose the Rasch model for our scaling procedure. For the Rasch model, we identified no misfitting items in either condition. All WMNSQs in the PBT and CBT fell below 1.10 and, thus, below our threshold for problematic misfit. Although the S-χ^2 tests identified a significant ($p > .05$) misfit of Item 20 for the PBT, the misfit was not considered severe after a visual inspection of the observed and expected item response curves. The other items in both conditions showed no significant misfit. The reliability estimates of .79 and .86 were good for PBT and CBT, respectively.

Identification of Anchor Items

Anchor items that exhibit invariant item parameters for the two assessment groups were identified by estimating a fully constrained multi-group item response model and evaluating the (weighted) root mean squared deviation (see Table 3). We selected Items 9, 13, and 21 with the lowest values of these indices. A comparison between the model with invariant item parameters for the three anchor items (*logLik* = -4128, Parameters = 44, AIC = 8378, BIC = 8518) and a model without constraints (*logLik* = -4127, Parameters = 46, AIC = 8382, BIC = 8528) showed that the constraints did not substantially impair model fit, $\Delta\chi^2(df = 2) = 1.48$, $p = .522$, thus, supporting the adopted invariance constraints. The latent mean of the CBT group was about -0.33 logits (Cohen's $d = -0.18$) smaller than the PBT group, suggesting lower average spelling competencies in the computer administration.

Analyses of Mode Effects

The absolute differences in item parameters between PBT and CBT were rather small (*Mdn* = 0.30, *Min* = 0.02, *Max* = 0.58). No item exhibited substantial differences greater than 0.63 (see Table 3). Moreover, the RMSD statistics fell between 0.00 and 0.06 in PBT (*Mdn* = 0.03) and between 0.00 and 0.10 in CBT (*Mdn* = 0.04). Only one item (Item 11) exhibited DIF greater than 0.08 for the CBT. However, the average DIF effect as given by the WRMSD

did not suggest pronounced overall DIF across both mode groups (see Table 1). Taken together, these results show that the item difficulties for most items were highly comparable between the two groups, and only a few items exhibited minor DIF. More importantly, these item-specific effects hardly distorted comparisons of the test scores. The test bias resulting from mode effects was about 0.16 score points (cDTF), 95% CI [-0.56, 0.90]. This means that, given the same spelling competencies, children working on a PBT are expected to achieve a test score that is about 0.16 points larger compared with children working on the same test as CBT. This translates into a percentage bias of less than 1% of the maximal test score of 22 points. Accordingly, the test for equivalence revealed that the expected test scores in the two mode conditions can be considered equivalent, $z = -1.91$, $p = .028$. These results were rather robust and replicated when limiting the bias analyses to the lower proficiency region between -3 and -1 on the logit scale (cDTF = 0.07, 95% CI [-0.83, 0.95]) or the higher proficiency region between 1 and 3 (cDTF = 0.25, 95% CI [-0.29, 0.84]).

The lack of substantial mode effects is also summarized in Figure 1. The left plot shows the expected test scores depending on the latent spelling competence for the PBT and CBT. Both test characteristic curves were substantially overlapped, thus, indicating highly similar measurement models for both mode groups. Consequently, the administration mode of the test also barely affected the expected proficiency distribution for samples with comparable spelling competence. The middle plot in Figure 1 shows that the respective distributions for PBT and CBT are largely indistinguishable, indicating that mode effects did not substantially affect the measured competencies. Finally, the marginal reliabilities as shown in the right plot of Figure 1 also emphasized comparable measurement precisions in both mode groups.

Discussion

Although paper-pencil assessments are still the *de facto* standard in many educational contexts, particularly for young children, digital procedures are increasingly adopted because they are often more economical, allow higher levels of standardization, and are more

motivating for children (e.g., Gnambs & Lenhard, 2023; Wright, 2020). To address the demands of educational practice for screening instruments that can be routinely administered in schools, the recently developed *SCHNAPP Spelling Test* in German (Schöfl et al., 2023) was created as a digital version to be administered on tablet with digital pens and as a traditional paper-based version. The findings from the current study demonstrated no substantial mode effects for the test. Rather, test scores obtained from the digital test can be considered equivalent to respective scores derived from the paper-based administration, conditional on the children's proficiencies. This underlines that digital testing does not necessarily have to affect cognitive measurements as long as the test procedures are highly comparable. For the *SCHNAPP Spelling Test*, the use of a digital pen guaranteed highly similar motoric demands as handwriting on paper. Moreover, in the digital test version children could work at their own pace and thus were not limited by either too slow or too quick item presentations. As a result, the *SCHNAPP Spelling Test* is one of rather few standardized competence tests in German for which equivalent paper- and computer versions are available (see also Gnambs & Lenhardt, 2023, for a German reading competence test). The availability of PBT and CBT versions of the test should facilitate its use in practice, for example, for large-scale screenings in school contexts.

Of course, some limitations need to be noted. For example, the presented analyses were limited to the internal structure of the test. But we did not study the validity of the test and whether mode effects might distort the prediction of later outcomes such as academic achievement. Similarly, we did not evaluate children's and teachers' perceptions of the different test versions and whether specific modes might be preferred. However, generally, reservations about digital testing are not to be expected (see Gnambs, 2022). After all, computers are found in most households and tablets are available in many families. Finally, future studies are encouraged that replicate these results with at-risk children, for example, students with special educational needs. Another aspect to analyze in future research with

bigger sample sizes are interaction effects for testing mode and characteristics such as migration background or familiarity with digital media. For the time being, the available evidence reported here suggests that the *SCHNAPP Spelling Test* can be used validly as PBT and CBT to capture the early spelling abilities of children in primary school.

References

Afonso, O. L., Martinez-Garcia, C., Cuetors, F., & Suarez-Coalla, P. (2020). The development of handwriting speed and its relationship with graphic speed and spelling. *Cognitive Development, 56*:100965. https://doi.org/10.1016/j.cogdev.2020.100965

Akaike, H. (1973). Information theory and extension of the maximum likelihood principle. on B. N. Petrov & B. F. Csáki (eds.), *Second International Symposium on Information Theory* (pp. 610–624). Akadémiai Kiadó.

Author, A. A. (2023). *Equivalence of computer- and paper-based administrations of the SCHNAPP spelling test in German for six-year-old children* [Computer code]. https://osf.io/6r4fh/

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (eds.), *Statistical Theories of Mental Test Scores,* (pp. 397–472). Addison-Wesley Publishing.

Bredel, U. (2015). Schriftspracherwerb. In U. Domahs & B. Primus (eds.), *Handbuch Laut, Gebärde, Buchstabe* [Handbook sound, gesture, letter] (pp.436-454). Walter de Gruyter.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Cheema, K., Fleming, C., Craig, J., Hodgetts, W. E., & Cummine, J. (2023). Reading and spelling profiles of adult poor readers: Phonological, orthographic and morphological considerations. *Dyslexia, 29*(2), 58-77. https://doi.org/10.1002/dys.1731

Caravolas, M. & Hulme, C., & Snowling, M. (2001). The foundations of spelling ability: Evidence from a 3 year longitudinal study. *Journal of Memory and Language, 45*, 751-774. https://doi.org/10.1006/jmla.2000.2785

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

Clinton V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading, 42*(2), 288–325. https://doi.org/10.1111/1467-9817.12269

De Sousa Lopes, M. C., & Bedulho, C. S. C. (2022). Relationship between reading fluency and spelling. *International Journal of Human Science Research, 2*(16), 1-11. https://doi.org/10.22533/at.ed.5582162220061

Domahs, F., de Bleser, R., & Eisenberg, P. (2001). Silbische Aspekte segmentalen Schreibens – neurolinguistische Evidenz [Syllabic aspects of segmental writing – evidence from neurolinguistics]. *Linguistische Berichte 185*, 13-29.

Domahs, F., Blessing, K., Kauschke, C., & Domahs, U. (2016). Bono Bo and Fla Mingo: Reflections of speech prosody in German second graders´ writing to dictation. *Frontiers in Psychology, 7:856*. https://doi.org/10.3389/fpsyg.2016.00856

Evertz, M. & Primus, B. (2013). The graphematic foot in English and German. *Writing Systems Research, 5*(1), 1-23. https://doi.org/10.1080/17586801.2013.765356

Fuhrhop, N., & Peters, J. (2013). *Einführung in die Phonologie und Graphematik* [Introduction into the phonology and graphemics]. J.B. Metzler.

Gnambs, T. (2022). The web-based assessment of mental speed: An experimental study of testing mode effects for the Trail-Making Test. *European Journal of Psychological Assessment*. Advance online publication. https://doi.org/10.1027/1015-5759/a000711

Gnambs, T., & Lenhard, W. (2023). Remote testing of reading comprehension in 8-year-old children: Mode and setting Effects. *Assessment*. Advance online publication. https://doi.org/10.1177/10731911231159369

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment, 5*(2).

Köhler, C., Robitzsch, A., & Hartig, J. (2020). A bias-corrected RMSD item fit statistic: An evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics*, *45*(3), 251–273. https://doi.org/10.3102/1076998619890566

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259-269. https://doi.org/10.1177/2515245918770963

Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology, 100,* 150-161. https://doi.org/10.1037/0022-0663.100.1.150

Li, Y. (2012). Examining the impact of drifted polytomous anchor items on test characteristic curve (TCC) linking and IRT true score equating. *ETS Research Report Series*, *2012*(1), 1–22. https://doi.org/10.1002/j.2333-8504.2012.tb02291.x

Lynch S. (2022). Adapting paper-based tests for computer administration: Lessons learned from 30 years of mode effects studies in education. *Practical Assessment, Research, and Evaluation, 27*:22. https://scholarworks.umass.edu/pare/vol27/iss1/22

Mesquita, A., Morais, I., Faísca, L., Reis, A., & Castro, S. (2022). Predictors of adult spelling in an orthography of intermediate depth. *Written Language & Literacy, 25*, 99-125. https://doi.org/10.1075/wll.00062.mes

Mombach, J., Rossi, F., Fernandes, D., & Soares, F. (2022). Children's impressions of early spelling assessment through handwriting on tablets vs. paper-based method. In ACM (ed) *Interaction Design and Children* (pp. 569-575). ACM. https://doi.org/10.1145/3501712.3535283

Neef, M. (2002). The reader´s view: sharpening in German. In M. Neef, A. Neijt & R. Sprout (eds.), *The relation of writing to spoken language* (pp. 169-191). Max Niemeyer.

Orlando, M., & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50–64. https://doi.org/10.1177/01466216000241003

Pan, S., Rickard, T., & Bjork, R. (2021). Does spelling still matter—and if so, how should it be taught? Perspectives from contemporary and historical research. *Educational Psychology Review, 33*, 1523-1552. https://doi.org/10.1007/s10648-021-09611-y

Penfield, R. D., & Camilli, G. (2006). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (eds.), *Handbook of Statistics* (Vol. 26, pp. 125-167). Elsevier.

Pohl, S., & Carstensen, C. H. (2013). Scaling the competence tests in the National Educational Panel Study. *Journal of Educational Research Online, 5*, 189–216. https://doi.org/10.1177/0013164414561785

R Core Team (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test analysis modules* [Computer software] (Version 4.1-4). https://CRAN.R-project.org/package=TAM

Robitzsch, A., & Lüdtke, O. (2020). A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psychological Test and Assessment Modeling, 62*(2), 233–279.

Schöfl, M., Steinmair, G., Zepnik, S., Zehetner, A., & Gnambs, T. (2023). Entwicklung eines digitalen Rechtschreibtests für die erste Klasse Grundschule: Dimensionalität und Reliabilität des SCHNAPP-Rechtschreibtests [Development of a digital spelling test for first grade: Dimensionality and reliability of the SCHNAPP spelling test]. *Lernen und Lernstörungen*. Advance online publication. https://doi.org/10.1024/2235-0977/a000404

Schroeder, S., Würzner, K. M., Heister, J., Geyken, A., & Kliegl, R. (2015). childLex–eine lexikalische Datenbank zur Schriftsprache für Kinder im Deutschen [childLex – a lexical database on literary language in German for children]. *Psychologische Rundschau, 66*(3), 155-165. https://doi.org/10.1026/0033-3042/a000275

Schwartz, G. E. (1978). Estimating the dimensions of a model. *Annals of Statistics, 6,* 461–464. https://doi.org/10.1214/aos/1176344136

Statistik Austria (2022). *Bildung in Zahlen 2020/21* [Education in numbers]. Statistik Austria.

White, S., Kim, Y. Y., Chen, J., & Liu, F. (2015). *Performance of fourth-grade students in the 2012 NAEP computer-based writing pilot assessment: Scores, text length, and use of editing tools* (No. NCES 2015-119). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/nationsreportcard/subject/writing/pdf/2015119.pdf

Wright A. J. (2020). Equivalence of remote, digital administration and traditional, in-person administration of the Wechsler Intelligence Scale for Children, (WISC-V). *Psychological Assessment, 32*(9), 809–817. https://doi.org/10.1037/pas0000939

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Zinn S., Landrock U., & Gnambs T. (2021). Web-based and mixed-mode cognitive large-scale assessments in higher education: An evaluation of selection bias, measurement bias, and prediction bias. *Behavior Research Methods, 53,* 1202-1217. https://doi.org/10.3758/s13428-020-01480-7

Open Science

We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact p values, effect sizes, and 95% confidence or credible intervals.

Open Data: We confirm that there is sufficient information for an independent researcher to reproduce all of the reported results (Author, 2023)

Open Materials: We confirm that there is sufficient information for an independent researcher to reproduce the reported methodology (Author, 2023).

OSF: https://osf.io/6r4fh/

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

Table 1

Sample Characteristics by Assessment Group

	Computer	Paper	Difference test
Sample size	148	242	
Number of schools/classes	6/11	9/14	
Percentage girls	45%	55%	$Chi^2(df = 1) = 2.92$, $p = .088$, $r_\varphi = .09$
Percentage home language only German	70%	80%	$Chi^2(df = 1) = 2.48$, $p = .116$, $r_\varphi = .08$
Percentage of parents with university degrees	71%	80%	$Chi^2(df = 1) = 3.48$, $p = .062$, $r_\varphi = .09$

Note. r_φ = Phi correlation.

Table 2

Comparison of Measurement Models

	logLik	Parameters	AIC	BIC	$\Delta\chi^2$	*df*	*p*
Paper-based spelling tests							
Rasch model	-2571	23	5196	5269			
Two-parametric model	-2550	44	5203	5342	42	21	.004
Computer-based spelling test							
Rasch model	-1556	23	3154	3226			
Two-parametric model	-1531	44	3142	3281	50	21	< .001

Note. $\Delta\chi^2$ = Log-likelihood ratio test comparing the two models; AIC = Sample-size adjusted Akaike's information criterion; BIC = Bayesian information criterion.

Table 3

Differential Item Functioning for Administration Mode

	Item difficulties (with 95% confidence interval)			Root mean squared deviation		WRMSD
Item	Computer	Paper	Difference	Computer	Paper	
1	-2.23 (-2.80, -1.67)	-2.67 (-3.14, -2.18)	0.43 (-0.31, 1.17)	0.03	0.04	0.04
2	-3.12 (-3.79, -2.45)	-3.32 (-3.92, -2.73)	0.20 (-0.69, 1.10)	0.05	0.02	0.03
3	-2.66 (-3.27, -2.05)	-2.41 (-2.86, -1.97)	-0.25 (-1.00, 0.51)	0.04	0.00	0.03
4	-2.80 (-3.42, -2.18)	-2.89 (-3.40, -2.38)	0.09 (-0.72, 0.90)	0.03	0.06	0.05
5	-1.77 (-2.30, -1.24)	-1.64 (-2.01, -1.27)	-0.13 (-0.78, 0.52)	0.02	0.03	0.03
6	-2.40 (-2.99, -1.82)	-2.46 (-2.91, -2.01)	0.06 (-0.68, 0.79)	0.03	0.04	0.04
7	-2.23 (-2.80, -1.67)	-2.12 (-2.53, -1.71)	-0.11 (-0.81, 0.69)	0.07	0.02	0.05
8	-1.12 (-1.63, -0.62)	-0.68 (-1.00, -0.35)	-0.47 (-1.05, 0.15)	0.08	0.04	0.06
9[#]	-1.36 (-1.67, -1.05)	-1.36 (-1.67, -1.05)	0.00	0.00	0.02	0.02
10	-0.73 (-1.23, -0.24)	-0.68 (-1.00, -0.35)	-0.06 (-0.65, 0.54)	0.03	0.02	0.03
11	-0.17 (-0.66, -0.32)	-0.54 (-0.86, -0.22)	0.38 (-0.21, 0.96)	0.10	0.04	0.07
12	-0.09 (-0.59, 0.40)	-0.46 (-0.78, -0.14)	0.36 (-0.22, 0.95)	0.05	0.05	0.05
13[#]	-0.29 (-0.57, -0.00)	-0.29 (-0.57, -0.00)	0.00	0.03	0.01	0.02
14	-2.18 (-2.74, 1.62)	-2.24 (-2.67, -1.81)	0.06 (-0.64, 0.77)	0.06	0.02	0.04
15	0.36 (0.14, 0.87)	0.35 (0.03, 0.66)	0.02 (-0.58, 0.61)	0.04	0.05	0.05
16	0.56 (0.06, 1.07)	0.27 (-0.05, 0.58)	0.30 (-0.30, 0.89)	0.02	0.04	0.04
17	1.63 (1.07, 2.20)	2.00 (1.60, 2.39)	-0.37 (-1.06, 0.33)	0.06	0.02	0.04
18	-0.32 (-0.81, 0.17)	0.27 (-0.05, 0.58)	-0.59 (-1.17, 0.00)	0.08	0.06	0.07
19	1.37 (0.82, 1.91)	0.86 (0.54, 1.19)	0.50 (-0.13, 1.14)	0.07	0.02	0.05
20	2.13 (1.51, 2.75)	1.59 (1.23, 1.95)	0.54 (-0.17, 1.26)	0.06	0.06	0.06
21[#]	0.17 (-0.12, 0.45)	0.17 (-0.12, 0.45)	0.00	0.03	0.00	0.03
22	2.13 (1.54, 2.75)	1.74 (1.37, 2.11)	0.40 (-0.33, 1.12)	0.04	0.00	0.03

Note. Difference = Difference in item difficulties with positive values indicating higher difficulty in computer-based assessment. *WRMSD* = Weighted root mean squared deviation. [#] Anchor item with equality constraint between groups.

Figure 1

Expected Total Scores and Reliability for Assessment Conditions