frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 24 July 2024
DOI 10.3389/feduc.2024.1376805

OPEN ACCESS

EDITED BY
Rüdiger Christoph Pryss,
Julius Maximilian University of Würzburg, Germany

REVIEWED BY
Young S. Seo,
University at Buffalo, United States
Michael Winter,
Julius Maximilian University of Würzburg, Germany

*CORRESPONDENCE
Susanne Seifert
✉ susanne.seifert@uni-graz.at

RECEIVED 26 January 2024
ACCEPTED 10 July 2024
PUBLISHED 24 July 2024

CITATION
Seifert S, Paleczek L, Schöfl M and Weber C (2024) Unveiling mode effects in grade 1 vocabulary assessment: the intriguing influence of test mode.
Front. Educ. 9:1376805.
doi: 10.3389/feduc.2024.1376805

Unveiling mode effects in grade 1 vocabulary assessment: the intriguing influence of test mode

Susanne Seifert[1,2]*, Lisa Paleczek[1,2], Martin Schöfl[3,4] and Christoph Weber[3,4]

[1]Department of Education Research and Teacher Education, University of Graz, Graz, Austria, [2]Research Center for Inclusive Education, Graz, Austria, [3]Department of Educational Sciences, University of Education Upper Austria, Linz, Upper Austria, Austria, [4]Research Institute for Developmental Medicine, Johannes Kepler University Linz, Linz, Upper Austria, Austria

Background: Vocabulary knowledge plays a pivotal role in academic development, particularly among Grade 1 students. To support students in their academic development, effective assessment instruments in educational settings are crucial. The GraWo (Graz Vocabulary Test) is introduced as a tool designed to evaluate receptive vocabulary in German-speaking countries in print and in digital mode.

Objectives: This study aims to investigate mode effects in the GraWo among Grade 1 students, comparing vocabulary gains in digital and print versions. Additionally, it explores the influence of student characteristics, such as gender and language status, and examines item-level differences between the two modes in order to gain a more comprehensive understanding of test performance.

Design: The research design entails a longitudinal approach, following children ($n = 421$) from the beginning to the end of Grade 1, varying the test modes (digital or print) only at second measurement (40% receiving the print version), while at first measurement all children worked with the digital version.

Results: Baseline comparisons of test mode groups indicated almost no significant differences. In terms of growth in vocabulary during Grade 1, an ANOVA with repeated measures revealed a main effect for time, indicating increased performance in both groups at second measurement. Moreover, an interaction effect between time and test mode group showed that the print group exhibited higher gains in the vocabulary test compared to the digital group. Further analysis using MNLFA confirmed that the print mode group outperformed the digital group overall and that four items were also individually affected by differences between the digital and print versions.

Conclusion: The study emphasizes the need for nuanced investigations into the impact of test mode on student performance and suggests incorporating observational methods to comprehensively understand student interactions with digital and print modes. In acknowledging potential variations in performance, educators and policymakers need to tailor practices to accommodate the demands of hybrid test procedures and to consider the role of digital competence in shaping testing experiences.

KEYWORDS

vocabulary, German, assessment, mode effect, test mode, grade 1

1 Introduction

Vocabulary plays a pivotal role in nurturing various linguistic abilities and is a crucial requirement for many skills essential in the academic context. Specifically, a child's proficiency in reading, including word recognition and comprehension, can be predicted by their vocabulary knowledge during their early years (Muter et al., 2004; Ennemoser et al., 2012; Juska-Bacher et al., 2021). The association between vocabulary and reading skills becomes even more pronounced among second language (L2) learners. Research consistently demonstrates that L2 learners tend to possess a lesser command of vocabulary in the language of instruction when compared to their first language (L1) counterparts (for example, Cremer and Schoonen, 2013; research specifically focusing on the German language: Klassert, 2011; Seifert et al., 2019). Consequently, L2 learners often exhibit poorer performance on reading assessments conducted in the language of instruction (Melby-Lervåg and Lervåg, 2014; Wendt and Schwippert, 2017). Hence, vocabulary knowledge is a relevant precursor for reading development, encompassing both receptive and productive vocabulary. Receptive vocabulary refers to the words that students can recognize and understand when they encounter them in listening or reading, while productive vocabulary includes the words they can use correctly in speaking and writing. Developing both types of vocabulary is crucial for Grade 1 students.

1.1 Vocabulary assessment

To foster vocabulary skills, reliable und valid assessment instruments for use in the school context are needed, as is also true regarding other precursor skills for written language (Ennemoser et al., 2012). Productive vocabulary, indicating a child's ability to express words, is usually assessed individually, especially when concerns about potential deficits arise. Conversely, receptive vocabulary, revealing a child's understanding of words, is intricately linked to reading proficiency since reading hinges on the comprehension of written words. Although it is possible to assess receptive vocabulary in a group setting with an entire class, there are only relatively few assessment tools available in German-speaking countries that support this setting.

The majority of German-language group assessments for receptive vocabulary skills primarily involve written language, such as requiring students to match one of four written words to a corresponding picture (e.g., word comprehension test as a subtest of the ELFE II reading comprehension test: Lenhard et al., 2020) or to decide whether a written word is a word of the German language or not (e.g., WOR-TE vocabulary test: Trautwein and Schroeder, 2019). These assessments tend to emphasize the orthographic component of vocabulary. However, according to Perfetti and Hart (2002), vocabulary encompasses two additional facets: the phonological component, encompassing knowledge about word pronunciation, and the semantic component, related to understanding word meanings. At the beginning of schooling, when reading skills are about to be acquired, methods that primarily focus on the orthographic component of vocabulary are not a suitable measurement technique. In contrast, the Graz Vocabulary Test (GraWo, Seifert et al., 2017) places particular emphasis on the semantic component of vocabulary by avoiding written language and therefore not relying on children's reading skills. In this receptive vocabulary assessment, designed for screening purposes and comprising only 30 items, children are tasked with matching a word presented orally to one of four pictures (see Figure 1). For the present study, we focus on Grade 1 students and assess receptive vocabulary knowledge.

1.2 Digital and print assessment

When using assessments in the school context, efficiency and user-friendliness are crucial. Group screenings like GraWo (Seifert et al., 2017 are preferred over complex individual tests. Additionally, digital assessment tools provide time-saving advantages. This increases their relevance and also furthers their acceptance (e.g., Neumann et al., 2019). Digital tools also enhance standardization of test instructions, item presentation, and response encoding, leading to more accurate and fairer assessments (Wang et al., 2021).

While the advantages of digital assessments are numerous, and their acceptance is widespread (e.g., Dawidowsky et al., 2021), the transition to digital assessments can only succeed if all schools are adequately equipped with digital devices, and if teachers as well as children possess the necessary expertise. Therefore, at least for a transitional period, both digital and traditional paper-based (print) assessments will coexist and remain relevant (Paleczek et al., 2021). Hybrid instruments that can be applied in both digital and print versions will continue to serve a purpose, allowing test users to flexibly utilize an instrument according to their individual needs. However, it is essential that the results from both versions are comparable since the goal is to assess the same skill with the instrument, and not to assess additional competencies such as digital skills (Puhan et al., 2007; Paleczek et al., 2021; Seifert and Paleczek, 2022).

1.3 Test mode effects

Due to the fundamental differences in the answering processes required for digital and print assessments, the research on mode effects continues to be a subject of interest among researchers (e.g., Wang et al., 2021). Test performance may also be influenced by the individual digital experience of examinees with the specific digital device that is used, which in turn raises concerns about potential variations in validity between digital and paper-based assessments. However, there is a lack of consistent research evidence regarding the equivalence and interchangeability of scores with respect to digital and print assessments. Across various instructional domains (reading, math, sciences), some studies (Johnson and Green, 2006; Puhan et al., 2007; Hamhuis et al., 2020) and meta-analyses (Wang et al., 2007, 2008) indicate comparability, while others suggest the opposite. In case of non-comparability, print assessments sometimes yield higher results compared to digital ones (Taherbhai et al., 2012; Lenhard et al., 2017; Backes and Cowan, 2019; Seifert and Paleczek, 2022; Wagner et al., 2022), and in other cases, digital assessments lead to higher scores (Lee et al., 2010; Wang et al., 2021).

The reasons for mode effects are still being debated. Some argue that mode effects are item-specific, tied to factors like response formats or item order (Buerger et al., 2019). Others suggest that differences in test-taking behavior and answering strategies explain such effects. For instance, computer-based test-takers tend to complete tests more quickly,

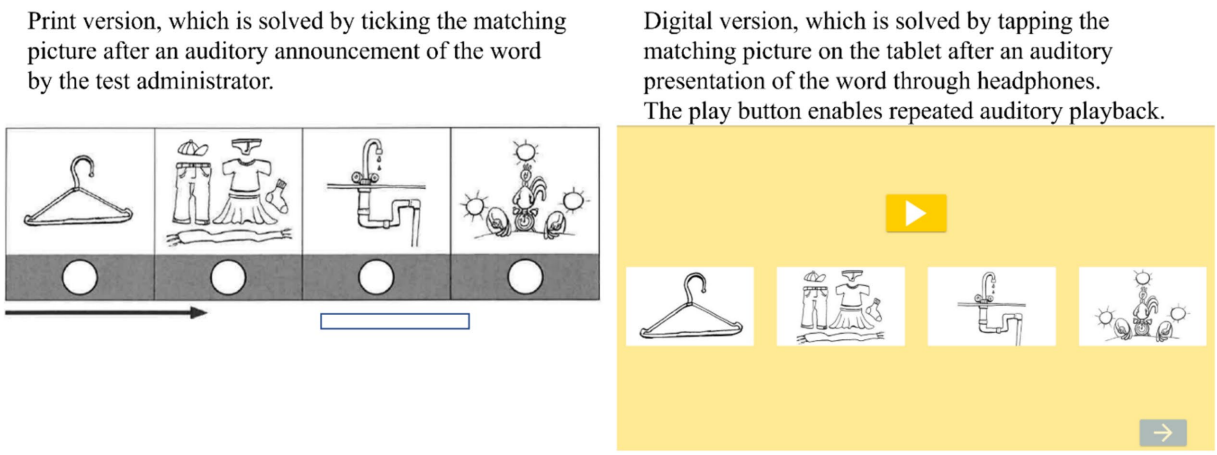

FIGURE 1
Print and digital versions of item "Kleidung" (clothes) in test GraWo.

guessing more often or responding more shallowly (Bodmann and Robinson, 2004; Karay et al., 2015; Lenhard et al., 2017; Singer-Trakham et al., 2019; Støle et al., 2020), although, here too, contradictory findings also exist (Steedle et al., 2022). Lenhard et al. (2017) noted that such behaviour is more prevalent in younger children (e.g., Grade 1 students) and when dealing with simpler tasks. Moreover, it is also prominent when a single test item is displayed on the screen (instead of more items or representation formats where scrolling is needed), as seen by Leeson (2006) as well as Bodmann and Robinson (2004). In contrast, paper-based assessments lead to more frequent answer reviews and amendments (Johnson and Green, 2006; Wang et al., 2021).

Mode effects have also been examined concerning specific student characteristics, with gender being a frequently explored variable. For instance, Jeong (2012) observed a difference between girls and boys, with girls performing significantly worse in digital assessments compared to paper-based ones in three out of four academic domains, while for boys this difference was only found in one domain. This was attributed to a gender gap in computer usage. However, most studies found no evidence of gender influencing mode effects (Clariana and Wallace, 2002; Poggio et al., 2005; Lee et al., 2010). Another student variable frequently discussed in relation to the mode effect is language status (L1 or L2). While some studies focused on English language learners (who can be called L2 learners), revealed mode effects in this specific sample (Hosseini et al., 2014), few have examined differences between L1 and L2 learners. Existing studies present a mixed picture. For example, Backes and Cowan (2019) identified stronger mode effects in language assessments for L2 learners compared to their L1 peers. In contrast, our own previous study (Seifert and Paleczek, 2022) found no moderating effect of language status on the mode effect found with a reading assessment tool. Apart from gender and language status, digital competence or familiarity with the corresponding digital devices seem to play a more significant role in moderating mode effects (Clariana and Wallace, 2002).

1.4 Scientific aim

The present study aims to contribute to a better understanding of mode effects, focusing on a vocabulary test previously employed to examine mode effects in the last year of kindergarten (Paleczek et al., 2021). While no significant differences in mean scores between the print and digital versions of this test were observed in the last year of kindergarten, a preference for the digital format was identified (as also revealed, e.g., by Gnambs and Lenhard, 2023). Furthermore, examiners were able to identify a distinct form of working behavior, indicating that children in the digital version worked more rapidly and with less self-monitoring.

The present study seeks to extend the investigation of mode effects with regard to the same instrument to the early stages of primary school, specifically, to explore whether the digital version truly aligns with the print version. Additionally, the study aims to scrutinize the moderating influence of the variables "language status" and "gender" concerning mode effects.

In detail, the present study examines mode effects in the use of the vocabulary test GraWo in Grade 1. Experimental design ensured that familiarity effects with the instrument and the vocabulary task itself were avoided. All children were given the digital version at the first measurement at the beginning of Grade 1 and the mode was only varied at the second measurement at the end of the school year.

The digital version was provided via tablets. We chose tablets over computers due to several factors. Tablets offer greater portability and ease of use, which are particularly important for young learners (Merchant, 2015). The touch-based interaction of tablets can be more intuitive and engaging for Grade 1 students, enhancing their learning experience (Ricoy and Sánchez-Martínez, 2020). Additionally, tablets are more commonly used in modern classrooms (Ricoy and Sánchez-Martínez, 2020) and households (Chaudron et al., 2017), making the study conditions more ecologically valid. Had we used computers, we might have encountered different challenges, such as the need for more advanced motor skills for mouse and keyboard use, potentially impacting the students' performance and the study's outcomes.

Based on our previous results with the same instrument in the last kindergarten year (Paleczek et al., 2021), we assume that there will be no significant differences between the two groups using different test versions at second measurement (digital vs. print). However, if there is a difference between the gains of the two groups, and both groups are comparable in terms of key baseline characteristics and composition - then it may be assumed that the test

version used at the second measurement does indeed have an impact on the results. We would then hypothesize that the group assessed with the paper version at second measurement would show higher gains (Taherbhai et al., 2012; Lenhard et al., 2017; Backes and Cowan, 2019; Seifert and Paleczek, 2022; Wagner et al., 2022). We also look at the influence of student gender and language status on mode effect. Based on our earlier work on mode effects in reading comprehension (Seifert and Paleczek, 2022), we suspect these two student characteristics will have no influence. However, given the mixed results in existing literature, it is crucial to empirically investigate these characteristics to ensure a comprehensive understanding of their potential impact in our specific context.

Furthermore, in addition to examining mode effects with respect to overall test scores, we also investigate mode differences in single items. In detail, we examine whether the probability of correctly solving an item differs between the two modes beyond what would be expected from mode differences on the overall test (i.e., differential item functioning; e.g., Bauer, 2023).

2 Materials and methods

This study is part of a longitudinal study concerning assessment of, and intervention in, reading and writing from the very beginning of schooling (see Schöfl et al., 2022).

In the present study, we followed children from the beginning (autumn 2021, Measurement 1) to the end of Grade 1 (summer 2022, Measurement 2). Vocabulary was assessed as part of a broader screening procedure which is described in Schöfl et al. (2022). At the beginning of Grade 1, vocabulary skills were assessed using the digital version of a vocabulary test provided on tablets. At the end of Grade 1, vocabulary was assessed using either the digital or the print version of the test (groups: digital vs. print). Only those students that performed the vocabulary test at both times were included in the present study.

The different measures used at first measurement in the screening procedure (assessing the domains sentence repetition and phonological information processing) are described below in section 2.2 (Instruments).

2.1 Participant recruitment and characteristics

The majority of the participants enrolled in this study were from a district in Upper Austria, encompassing four prominent community-based schools. Initial contact and invitations for study participation were extended to eligible schools through telephone communication, followed by subsequent in-person visits. All school headmasters provided their consent for participation, and an additional four schools expressed interest and subsequently joined the project. Ultimately, parents of 459 students agreed (providing written permission) at the beginning of Grade 1 that their child(ren) could participate in the study. Data on the GraWo vocabulary test is available solely at Measurement 1 for 38 students [8.28% of the consented sample; 20 of them male, 30 monolinguals in German (language status L1)]. For the remaining 421 children (91.72% of the consented sample), data existed for both measurements (beginning and end of Grade 1). Consequently, these students were included in the analyses for this study.

The resulting study sample exhibited a diverse mix of children, and mirrored the characteristics of the Austrian primary school population in terms of gender, the proportion of L2 learners, and parental educational levels.

The characteristics of participants are representative for Austrian elementary school children. Of the 421 students, 48.9% ($n = 206$) are female, 73.6% ($n = 310$) are monolinguals in German (language status L1). These proportions correspond to those found in Austrian school statistics (Statistik Austria, 2022). Parents' highest educational attainment was used as a proxy for the children's socioeconomic background. The sample consisted of students with parents from all educational backgrounds: among the mothers, 7.6% had a maximum educational attainment of a secondary school diploma, 28.3% completed an apprenticeship or a VET school, 19.2% had a high school diploma, and 38.2% had a university diploma. The fathers' educational levels were comparable (7.8% secondary school diploma, 34% apprenticeship or VET school,14.3% high school diploma, and 30.9% with a university degree).

The individualized screening process commenced in the autumn of 2020, and was initiated within 2 weeks of school onset in 27 classes. Within 3 weeks, assessments had been completed for 85% of the sample. Subsequently, over the following 2 weeks, children who were either unwell or unable to attend during the initial survey period were also assessed.

2.2 Instruments

The Graz Vocabulary Test (GraWo; Seifert et al., 2017) was used to assess vocabulary skills. The GraWo is a standardized screening instrument that assesses receptive vocabulary in first to third graders. It consists of two sample items and 30 test items (5 verbs, 5 adjectives, 5 prepositions, and 15 nouns, with the latter being split into 5 monomorphematic nouns, 5 composite nouns and 5 nouns referring to categories; for a list of all items, see Supplementary Table S1). The children are required to select one out of four pictures matching a word that is presented audibly. This test can be administered in various formats and settings, broadly classified as group or single settings, and either as a print version or a digital version (conducted on tablets). For the scope of our study, we provided the print version in a single setting, where the word was pronounced out loud by the instructor. The digital version was delivered in a single setting as well, in which children used headphones to listen to prerecorded pronunciations by professional speakers (as illustrated in Figure 1). This approach was chosen to ensure uniformity in the conditions of test administration across both modes, allowing for a fair comparison of performance outcomes.

This vocabulary test was either provided in the digital version [as a part of a wider screening procedure, for all of the students at the beginning of Grade 1 ($N = 421$) and for nearly 60% of the sample at the end of Grade 1 ($n = 249$)], or as a print version [for nearly 40% of the sample at the end of the school year ($n = 172$)]. Reliability data are given for the print version of the GraWo (Seifert et al., 2017): Cronbach's Alpha ranged from 0.89 (end of Grade 1) to 0.82 (end of Grade 2). Retest reliability was $r_{tt} = 0.93$ (Grade 1).

A parent questionnaire was used to obtain information about children's L1 and socioeconomic background. When this indicated that the L1 was only German, or that contact with German occurred from birth up to and including the age of 2, then children were classified as having language status L1. Children whose contact with German occurred only after the age of 2 were classified as having language status L2.

Additionally, a screening of potential precursors of reading skills (covering sentence repetition: adapted from Ibrahim et al., 2018; phonological awareness (PA): Schöfl et al., 2022; rapid automatized naming (RAN) objects: Schöfl et al., 2022; RAN digits: Denckla and Rudel, 1974; letter knowledge: Schöfl et al., 2022; phonological working memory: Grob et al., 2009) was used at the beginning of Grade 1 (for more information, see also Schöfl et al., 2022). The results of these tests were used to analyze whether the two test mode groups differed in terms of baseline results.

2.3 Procedure

Before assessment, school principals received information and a letter for parents explaining the testing procedure, along with consent forms, a data protection declaration, and questions about children's language and parents' educational background. Teachers entered student names into an online database, converting them to IDs for tablet use. Pre-service teachers enrolled in a university seminar on testing procedures carried out the testing and received course credit for it (amounting to 4 h). The materials for testing were brought to the schools by a research coordinator. On test mornings, the test team (pre-service teachers and project staff) selected students alphabetically, and assessed them individually. Instructor and child were seated at a table across from each other. Each child was given a tablet and guided through the screening process, starting with an introduction to the friendly dragon SCHWUPP. App navigation allowed independent use, with instructors intervening if needed. Instructions were recorded as audio files, opened automatically, and could be listened to repeatedly if necessary. The assessment, including all subtests, averaged 38.4 min per child (SD = 9.3). Vocabulary assessment with the GraWo was just one part of the screening.

At the end of the school year, the procedure remained similar, except for one key modification: to address the present research questions, classes were randomly assigned to either receive the digital mode again or switch to the print mode of the vocabulary test. Half of the classes, comprising 40% of the students, received the print version of the vocabulary test instead of the digital one. Due to logistical practicalities, such as the allocation of tablets and paper materials, once a class was randomly selected for a test mode, all students within that class were tested in the same mode to maintain consistency and manage resources efficiently.

The total number of items solved was either recorded directly by the app in the case of digital testing, or was counted manually by the research team. Data on individual items are available for all children in the digital version for both measurements (measurement 1: $n = 421$ of 421 students, measurement 2: $n = 249$ of 249 students, 100%), but only for some of the children in the print version ($n = 69$ of 197 students, 35%). Our analysis revealed that the two print version groups, one with available data on individual items ($n = 69$) and one without ($n = 128$), did not differ significantly in their performance on the GraWo print version of the test [$t(195) = 1.37$, $p = 0.17$] and were comparable in gender distribution [$\chi^2(1) = 0.64$, $p = 0.424$, $\varphi = -0.06$] and distribution of L1 and L2 students [$\chi^2(1) = 0.98$, $p = 0.366$, $\varphi = 0.07$]. However, disparities were noted in terms of socioeconomic background: the group with available data had a lower socioeconomic status, as suggested by mothers' and fathers' highest educational attainment (mothers': $U = 2894.00$, $Z = -3.40$, $p < 0.001$; fathers': $U = 3032.50$, $Z = -2.69$, $p = 0.007$), and had fewer years of kindergarten attendance ($U = 3561.50$, $Z = -2.05$, $p = 0.040$), than the group lacking individual item data.2.5 Methods.

First, the two test mode groups (digital vs. print) were analyzed in terms of sample composition (using Chi-square tests) and baseline differences (using t-tests).

Second, to analyze for mode effects, an ANOVA with repeated measures with the inner-subject variables "GraWo scores beginning of Grade 1" and "GraWo scores end of Grade 1" and the inter-subject factor test mode group (digital at the end of Grade 1 vs. print at the end of Grade 1) was conducted to find out whether it makes a difference whether the GraWo was used digitally or in print at the end of Grade 1. A significant interaction of time and mode would indicate a mode effect. Additionally, the inter-subject factors gender (girls vs. boys) and language status (L1 vs. L2) were included to look for interaction effects with the test mode group.[1]

Third, to deepen the analyses on mode effects, we applied moderated non-linear factor analysis (MNLFA; Bauer, 2023) – a relatively new and flexible approach – to test for differential item functioning (DIF). We used the aMNLFA R-package (Gottfredson et al., 2019) together with Mplus 8 (Muthén and Muthén, 1998-2017). The DIF analysis consisted of the following steps (see Gottfredson et al., 2019, p. 68): (1) We tested whether a unidimensional model for the GraWo fit the data well, using the following guidelines for the evaluation of model fit (Schermelleh-Engel et al., 2003): $\chi^2/df \leq 2$, comparative fit index (CFI) ≥ 0.975, and root mean square error of approximation (RMSEA) ≤ 0.05 for a good fit and $\chi^2/df \leq 2$, CFI ≥ 0.975, RMSEA ≤ 0.05 for an acceptable fit. As the standardizes root mean square residual (SRMR) has been shown to over-reject models with binary indicators, we do not report this index (Yu, 2002). At this stage we used a weighted least square parameter estimation (WLSMV) with a probit link, as this estimation provides more model fit indices than maximum likelihood (ML) estimation. For all subsequent analyses, a ML estimation with Monte Carlo integration was required. (2) We assessed whether there are mode effects on the GraWo latent mean and the GraWo latent variance. At this stage, an α-level of 0.10 was used. (3) We applied an item-by-item approach to test for DIF, i.e., we tested mode effects on a single item (threshold and loading) while holding thresholds and loadings of all other items constant. At this stage, due to multiple testing issues an α-level of 0.05 was used. (4) Finally, we tested all significant mode effects of step 2 and 3 in a simultaneous model. The Benjamini-Hochberg correction was applied to correct for the effects of multiple testing. Mode effects that remained statistically significant in this final simultaneous model

1 Taking a different approach on the analysis of mode effects, we also conducted a regression analysis with GraWo at the end of Grade 1 as dependent variable and mode as independent variable. We additionally controlled for GraWo scores assessed at the beginning of Grade 1, gender and language status.

are reported below. For more details on the DIF testing procedure see Gottfredson et al. (2019).

3 Results

3.1 Baseline comparison of the test mode groups

With respect to sample composition, we investigated whether the proportion of girls and the proportion of students with language status L1 was comparable in the test mode groups. Results show neither a significant relationship between group and gender [$\chi^2(1) = 3.81$, $p = 0.051$, $\varphi = 0.09$], nor between group and language status [$\chi^2(1) = 0.36$, $p = 0.548$, $\varphi = 0.03$].

As shown by the Mann–Whitney-*U*-Test, socioeconomic background was not comparable in the test mode groups. The distributions of parents' highest educational attainment differed between both groups. There was a statistically significant difference between test mode group and mothers' highest educational attainment ($U = 17382.50$, $Z = -2.30$, $p = 0.022$) and fathers' highest educational attainment ($U = 16763.00$, $Z = -2.58$, $p = 0.010$), respectively. The digital test mode group showed higher educational attainments of both mothers and fathers than the print test mode group.

On considering the number of years of kindergarten attendance, no significant difference was found between the two groups ($U = 19920.50$, $Z = -0.56$, $p = 0.577$).

In terms of baseline differences, we first investigated the test mode groups with regard to differences in the mean values of the vocabulary test at the beginning of Grade 1. Results show no significant differences at first measurement in GraWo [$t(419) = 1.16$, $p = 0.249$; for mean values, see also Table 1]. Moreover, no other baseline assessments revealed differences between the two groups either, with one exception: the digital group achieved significantly higher sentence repetition scores (see Table 1). To facilitate a clear understanding of this particular outcome, it is pertinent to expound upon the sentence repetition task. This assessment comprised a block of 15 items representing morphosyntactic constructions with varying degrees of complexity. Each item was scored according to whether or not the sentence was reproduced correctly. Correct reproduction required that the sentence structure be mirrored precisely as presented, irrespective of any articulatory deficits that did not interfere with structural accuracy (for more information, see also Schöfl et al., 2022).

To summarize, there are hardly any differences between the two test mode groups. The existing differences (highest educational attainments of parents and sentence repetition) favor the group that was tested digitally at the second measurement compared to the group tested in print.

3.2 Comparison of vocabulary growth during the first grade across the test mode groups

Table 2 presents the means of the scores in the vocabulary test GraWo for both test mode groups.

An ANOVA with repeated measures showed a main effect for time [$F(1, 419) = 377.93$, $p < 0.001$, $\eta^2 = 0.47$], showing that both groups solved more items at the end of Grade 1 compared to the beginning of Grade 1. Additionally, an interaction effect of time and test mode group was uncovered [$F(1, 419) = 27.34$, $p < 0.001$, $\eta^2 = 0.06$]. The test mode group "print" showed higher gains in the vocabulary test (see Figure 1), suggesting a positive print mode effect[2] (Figure 2).

No main effects for gender [$F(1, 413) = 0.00$, $p = 0.996$] and language [$F(1, 413) = 0.00$, $p = 0.959$] were detected, nor were any further interaction effects with the test mode group revealed.

3.3 Comparison of test modes through DIF analyses

Before performing DIF analyses, we fitted a one-factor model to the data to evaluate whether the GraWo is unidimensional. In detail, we fit a one-factor model with loadings constrained to be equal (corresponding to a one-parameter (1pl) item response theory (IRT) model) and a one-factor model with loadings freely estimated [corresponding to a two-parameter (2pl) IRT model].

The 2pl model resulted in a good fit ($\chi^2(405) = 454.8$, $p < 0.05$; CFI = 0.980; RMSEA = 0.018, 90%-CI [0.023, 0.026]), whereas, the more restrictive 1pl model did not fit the data well ($\chi^2(434) = 802.5$, $p < 0.001$; CFI = 0.855; RMSEA = 0.047, 90%-CI [0.042, 0.052]). The 2pl model was also supported by the results of a χ^2-difference test ($\chi^2(29) = 171.0$; $p < 0.001$). Taking the 2pl as a starting point, we subsequently tested DIF using MNLFA. Although the digital mode group demonstrated more consistent performance, as evidenced by a significantly ($p < 0.001$) smaller standard deviation ($SD = 0.87$ while the SD in the print mode group was fixed at 1), it was the print mode group that, on average, outperformed the digital mode group in the vocabulary test scores (latent mean difference $d = 0.400$, $p < 0.001$). Finally, four items showed statistically significant DIF. Items 8, 15, and 21 were less often correctly solved in the digital mode group than in the print group compared to what would be expected from the overall mode effect. In contrast, the opposite was true for item 14, i.e., children were more likely to solve this item when it was offered in digital mode.

4 Discussion

4.1 Test mode effect in vocabulary assessment

The findings of this study stress the significant impact of test mode on student scores in assessments in Grade 1 students, particularly in the realm of vocabulary testing. Despite both test mode groups displaying comparable vocabulary results at the initial measurement, where both were assessed digitally, the subsequent gains exhibited noteworthy differences depending on test mode. Notably, students

2 This conclusion was also supported by the results of a supplementary regression analysis. Controlling for GraWo at the beginning of Grade 1, gender, and language status, students from the print mode group scored 1.65 points higher (b=1.65, SE=0.31, p<0.001, Cohen's d=0.36) on the GraWo at the end of Grade 1.

TABLE 1 Domains assessed at the beginning of the school year, information about the used assessments and results of the two test mode groups and tests of mean differences.

Domain[a]	Subtests	Number of test items	Test medium	Measured value	Results of both groups *M (SD)*		Differences between groups[a], *df* = 419
					Digital	Print	
Vocabulary		30	Digital	Score	21.17 (4.82)	20.58 (5.56)	$T = 1.16$, $p = 0.249$
Sentence repetition		15	Digital	Score	10.32 (3.77)	9.24 (4.25)	$T = 2.74$, $p = 0.006$
Phonological information processing	PA: Rhyme detection	10	Digital	Score	8.91 (1.25)	8.67 (1.42)	$F = 3.32$, $p = 0.07$
	PA: Syllable count	10	Digital	Score	8.78 (1.87)	8.71 (1.75)	$F = 0.15$, $p = 0.70$
	PA: Initial phoneme detection	10	Digital	Score	8.53 (1.67)	8.44 (1.97)	$F = 0.30$, $p = 0.58$
	RAN objects	30	Print	Time in seconds	34.03 (8.75)	34.80 (9.59)	$F = 0.72$, $p = 0.40$
	RAN digits	30	Print	Time in seconds	33.92 (11.97)	34.07 (16.53)	$F = 0.01$, $p = 0.91$
	Letter knowledge	26	Print	Score	14.76 (8.16)	15.31 (8.26)	$F = 0.46$, $p = 0.50$
	Phonological working memory – letter number span forward	adaptive	Verbal (instructor)	Score	7.32 (1.84)	7.32 (2.44)	$F = 0.00$, $p = 0.97$
	Phonological working memory – letter number span backward	adaptive	Verbal (instructor)	Score	3.90 (1.96)	4.10 (2.18)	$F = 0.96$, $p = 0.33$

[a]Domains without subtests: *t*-tests were conducted, Domains with at least two subtests: multivariate variance analyses were conducted.

TABLE 2 GraWo scores for both test mode groups.

Test mode group	Beginning of Grade 1 (measurement 1) – all students digital		End of Grade 1 (measurement 2)	
	M	*SD*	*M*	*SD*
Digital	21.17	4.82	23.71	4.62
Print	20.58	5.56	24.99	4.40

assessed with the print version at the second measurement demonstrated significantly higher gains compared to their counterparts who were tested using the digital version. Assessment of DIF showed that the mode effect occurs overall, but that four items are also individually affected by these differences between the digital and print versions beyond what we would expect from the overall mode effect alone. For item 8, one possible explanation for the advantage of the print version could be that it is easier to differentiate from an auditory distractor (target item "hell" [bright], distractor "Fell" [fur], see also Supplementary Table S1) due to the mouth of the instructor being visible in the print version. Such cues are absent in the digital version. For the other three items, we cannot explain the differences between the print and digital modes. Future studies are needed to address this topic more systematically.

The revealed test mode differences were contrary to our original assumptions (i.e., based on our previous results with the same instrument in the last kindergarten year no differences were found) (Paleczek et al., 2021). However, the results are in line with the existing literature, where higher student scores are found for print version than for the digital version assessment (e.g., Taherbhai et al., 2012; Lenhard et al., 2017; Backes and Cowan, 2019; Seifert and Paleczek, 2022; Wagner et al., 2022). The present study even adds further weight to the argument, since at second measurement, all students were familiar not only with the assessment instrument, but also with the digital test mode as all of them had been assessed with the digital version at the first measurement (at the beginning of Grade 1). They knew what to do in the digital mode. Hence, inexperience with the digital instrument cannot be used as the cause of the mode effect. Moreover, the rare baseline differences we detected between the two groups actually favored the digital group (highest parental educational attainment, sentence repetition). Yet, apparently, these were not sufficiently strong to counteract the influence of test mode.

As evidenced by prior studies (e.g., Lenhard et al., 2017), it is likely that the unique demands present in digital assessment instruments contribute to the observed differences. In the GraWo, both the setting and the response format differ for the digital and print versions. The digital setting, with instructions received via headphones, is rather individualized and self-directed, allowing students to decide for themselves when to proceed. In contrast, the print setting is a guided process, where students only proceed when the instructor decides to do so. The specific response format also differs between the digital and the print mode. In the digital version, students tap the correct image and the "proceed" button on a tablet, which makes items visible individually, one after the other. Although listening to the audio

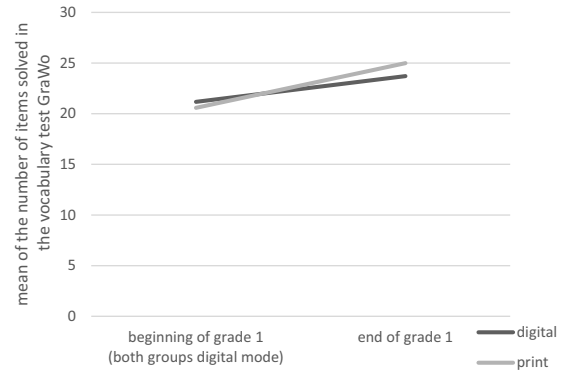

FIGURE 2
GraWo scores at the beginning and end of Grade 1, using either the digital or print version of the GraWo at the end of Grade 1.

announcement again and returning to a previous item is possible, students may not make sufficient use of these functions. In the print mode, students make a cross in the test booklet following the instructor's command. The instructor can easily be asked to repeat the announcement. Additionally, each page in the print version presents five items, which leaves previous items visible and might facilitate access to rethinking and, if necessary, correction of earlier answers. Observations on students' test behaviors with the kindergarten version of this test (Paleczek et al., 2021) suggest that students using the digital version in the present study provided faster and potentially less-considered responses, as has been found in previous studies (e.g., Bodmann and Robinson, 2004; Lenhard et al., 2017; Singer-Trakham et al., 2019). Previous studies have also shown this effect to be especially prominent when using digital assessments with single items (and not multiple ones) displayed on the screen (Bodmann and Robinson, 2004; Leeson, 2006), as was the case in the present study. This highlights the potential impact of the self-directed nature of the digital mode of the GraWo, where students may aim to complete tasks more swiftly and perhaps be less accurate in their responses.

Despite these insights, there remains a dearth of observational studies comparing the actual behaviors of students in digital and print modes, especially in the investigation of mode effects. Future research should prioritize such observations in order to gain a more nuanced understanding of how students interact with different test modes and also to shed more light on the reasons for different student interactions.

In the current study, we also delved into student characteristics concerning the mode effect. The findings revealed that neither gender nor language status exerted an influence on the mode effect, which is also in line with previous research (for gender: Clariana and Wallace, 2002; Poggio et al., 2005; Lee et al., 2010; for language status: Seifert and Paleczek, 2022). However, a significant factor remained largely unexplored: digital competence. This omission is particularly noteworthy given the evolving landscape of technology in education. Existing research (e.g., Clariana and Wallace, 2002) emphasizes the importance of considering digital competence in understanding students' interactions with digital and print modes. However, while frequency of device use may not be the primary determinant, the actual application skills, such as tablet orientation and proficiency in app-specific functions (e.g., tapping), may moderate effects. Future studies investigating mode effects need to scrutinize digital competence and associated application components in order to further elucidate the nuanced dynamics at play.

In conclusion, this study contributes valuable insights into the multifaceted impact of test mode on student scores, emphasizing the need for a comprehensive understanding of setting, task requirements, and student digital competence. The findings prompt a re-evaluation of assumptions about digital testing behaviors and underscore the imperative for more extensive observational research in this domain.

4.2 Limitations

This study shares common limitations with previous research (e.g., Gnambs and Lenhard, 2023) as we did not employ a true experimental design that involves randomly assigning children to test mode groups. Instead, for administrative convenience, entire classrooms were assigned to either the digital or print condition at the second measurement. While this approach was logistically more feasible, it introduced a limitation in terms of establishing a more robust causal relationship due to the potential for selection bias. Specifically, differences in classroom environments, teachers' instructional styles, or other classroom-level characteristics could systematically differ between conditions and influence student performance. To address this, we have controlled for baseline performance and other individual-level characteristics, such as gender and language status. This strategy aimed to reduce the effects of measured confounders at the student level. We acknowledge, however, that given the low number of classrooms that were randomly assigned to the test mode groups, unmeasured confounders (at student and classroom level) that differ by chance between print and digital mode classrooms may be an issue. The lack of randomization at individual level limits the strength of causal inferences that can be drawn from our findings. Therefore, while our analysis provides valuable insights into the test mode effect, the results should be interpreted with an understanding of these methodological constraints. Going forward, future studies may benefit from employing design improvements such as randomization at individual level to enhance causal inference. Replication in diverse educational settings with random assignment would further validate our findings and contribute to a more comprehensive understanding of the impact of assessment modes.

The fact that the processing of the individual items in the print version was not available for all children at second measurement, but only for around 35% of the children, represents a further limitation. It was shown that the GraWo scores did not differ between those children for whom the data for the processing of the individual items was available and those for whom this was not the case. However, there were observed differences in socioeconomic background and years of kindergarten attendance between the group of children in the print version with individual item data and the group without such data. These factors are known to influence educational outcomes, and as such, they may have impacted our results. The group with complete data had less favorable socioeconomic conditions and fewer years of kindergarten attendance, which could potentially confound the comparisons of performance between the print and digital groups as well. Future research should aim to ensure a more homogeneous sample or control for these variables to accurately isolate the effects of the mode of test administration.

While our research sheds light on the comparative efficacy of print versus digital test modes, we must note the omission of participants' digital competence data and the absence of observational data as two notable limitations. The potential variation in digital competence among students could be a confounding factor that influences their ability to perform optimally in a digital environment. Simultaneously, without observational data, the behavioral and interactive facets of testing that might differentially impact performance in digital versus print formats remain unexplored. Future studies should seek to incorporate both assessments of digital competence and structured observations during test-taking to derive a more granular understanding of how these variables intersect to influence students' assessment outcomes. Such an approach would contribute to a more rigorous and nuanced understanding of the dynamics involved in mode effects.

While our findings contribute valuable insights into the efficacy of print versus digital test modes, we acknowledge that the generalizability of these results may be limited by contextual factors unique to our study's setting. The implications of this research could vary significantly in different educational environments, particularly across regions with disparate levels of technological infrastructure and varying pedagogical traditions. For instance, students in areas with limited access to digital devices or those not routinely integrated into classroom learning may respond differently to test modes than those in our study cohort. Additionally, the familiarity with specific test formats due to prevailing instructional strategies could impact the outcomes observed in other contexts. Therefore, we emphasize the necessity for caution when extrapolating our study's conclusions to dissimilar settings. To ascertain the broader applicability of these findings, we endorse the undertaking of replication studies encompassing a diverse array of educational contexts. This would enhance the understanding of the conditions under which the observed advantages of either test mode are most pronounced.

4.3 Implications for research and practice

The findings emphasize the need for further investigations into the impact of test mode on student scores in hybrid assessment tools, considering not only test outcomes but also the cognitive processes and behaviors associated with different modes. In practice, educators and policymakers should acknowledge the potential differences in student scoring arising from test mode variations. The study highlights the importance of considering the specific demands of hybrid test procedures and tailoring educational practices accordingly. Moreover, educators need to be aware of the potential influence of digital competence on students' assessment experiences. As technology continues to play a central role in education, understanding and addressing the nuances brought forth by different test modes are imperative for informed decision-making in both research and educational settings.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/y29ps/?view_only=a8b09d29c9bc4509aa4a84a60f2e91e9.

Ethics statement

The studies involving humans were approved by Regional School Board for Upper Austria (Bildungsdirektion Upper Austria). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

SS: Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. LP: Writing – review & editing. MS: Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Supervision, Writing – review & editing. CW: Formal analysis, Methodology, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. All costs related to this research project were borne by the University of Education Upper Austria and the Research Institute for Developmental Medicine (RID), Johannes Kepler University Linz. The tablets were provided by the project team with funds from the University of Education Upper Austria. The article processing charge was covered by the University of Graz Open Access Publishing Fund.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2024.1376805/full#supplementary-material

References

Backes, B., and Cowan, J. (2019). Is the pen mightier than the keyboard? The effect of online testing on measured student achievement. *Econ. Educ. Rev.* 68, 89–103. doi: 10.1016/j.econedurev.2018.12.007

Bauer, D. J. (2023). Enhancing measurement validity in diverse populations: modern approaches to evaluating differential item functioning. *Br. J. Math. Stat. Psychol.* 76, 435–461. doi: 10.1111/bmsp.12316

Bodmann, S. M., and Robinson, D. H. (2004). Speed and performance differences among computer-based and pencil-paper tests. *J. Educ. Comput. Res.* 31, 51–60. doi: 10.2190/GRQQ-YT0F-7LKB-F033

Buerger, S., Kroehne, U., Koehler, C., and Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Stud. Educ. Eval.* 62, 1–9. doi: 10.1016/j.stueduc.2019.04.005

Chaudron, S., Di Gioia, R., and Gemo, M. (2017). Young children (0-8) and digital technology - a qualitative study across Europe. Luxembourg: Publications Office of the European Union.

Clariana, R., and Wallace, P. (2002). Paper–based versus computer–based assessment: key factors associated with the test mode effect. *Br. J. Educ. Technol.* 33, 593–602. doi: 10.1111/1467-8535.00294

Cremer, M., and Schoonen, R. (2013). The role of accessibility of semantic word knowledge in monolingual and bilingual fifth-grade reading. *Appl. Psycholinguist.* 34, 1195–1217. doi: 10.1017/S0142716412000203

Dawidowsky, K., Holz, H., Schwerter, J., Pieronczyk, I., and Meurers, D. (2021). Development and evaluation of a tablet-based Reading fluency test for primary school children, 1–17. doi: 10.1145/3447526.3472033

Denckla, M. B., and Rudel, R. (1974). Rapid "automatized" naming of pictured objects, colors, letters and numbers by normal children. *Cortex* 10, 186–202. doi: 10.1016/S0010-9452(74)80009-2

Ennemoser, M., Marx, P., Weber, J., and Schneider, W. (2012). Spezifische Vorläuferfertigkeiten der Lesegeschwindigkeit, des Leseverständnisses und des Rechtschreibens. *Zeitschrift Entwicklungspsychologie Pädagogische Psychologie* 44, 53–67. doi: 10.1026/0049-8637/a000057

Gnambs, T., and Lenhard, W. (2023). Remote testing of Reading comprehension in 8-year-old children: mode and setting effects. *Assessment* 31, 248–262. doi: 10.1177/10731911231159369

Gottfredson, N. C., Cole, V. T., Giordano, M. L., Bauer, D. J., Hussong, A. M., and Ennett, S. T. (2019). Simplifying the implementation of modern scale scoring methods with an automated R package: automated moderated nonlinear factor analysis (aMNLFA). *Addict. Behav.* 94, 65–73. doi: 10.1016/j.addbeh.2018.10.031

Grob, A., Meyer, C. S., and Hagmann-von Arx, P. (2009). Intelligence and Development Scales (IDS).: Intelligenz-und Entwicklungsskalen für Kinder von 5–10 Jahren. Bern: Verlag Hans Huber.

Hamhuis, E., Glas, C., and Meelissen, M. (2020). Tablet assessment in primary education: are there performance differences between TIMSS' paper-and-pencil test and tablet test among Dutch grade-four students? *Br. J. Educ. Technol.* 51, 2340–2358. doi: 10.1111/bjet.12914

Hosseini, M., Abidin, M. J. Z., and Baghdarnia, M. (2014). Comparability of test results of computer based tests (CBT) and paper and pencil tests (PPT) among English language learners in Iran. *Procedia Soc. Behav. Sci.* 98, 659–667. doi: 10.1016/j.sbspro.2014.03.465

Ibrahim, L., Hamann, C., and Öwerdick, D. (2018). Identifying specific language impairment (SLI) across different bilingual populations: German sentence repetition task (SRT), vol. 2. Proceedings of the 42nd annual Boston University Conference on Language Development, 1–14. Somerville, MA: Cascadilla Press.

Jeong, H. (2012). A comparison of the influence of electronic books and paper books on reading comprehension, eye fatigue, and perception. *Electron. Libr.* 30, 390–408. doi: 10.1108/02640471211241663

Johnson, M., and Green, S. (2006). Online mathematics assessment: the impact of mode on performance and question answering strategies. *J. Technol. Learn. Assesment* 4, 1–33.

Juska-Bacher, B., Röthlisberger, M., Brugger, L., and Zangger, C. (2021). "Lesen im 1. Schuljahr: Die Bedeutung von phonologischer Bewusstheit, Benennungsgeschwindigkeit und Wortschatz" in Weiterführende Grundlagenforschung in Lesedidaktik und Leseförderung: Theorie, Empirie, Anwendung. eds. S. Gailberger and C. Sappok, (Bochum: Universitätsbibliothek der Ruhr Universität Bochum) 11–26.

Karay, Y., Schauber, S. K., Stosch, C., and Schüttpelz-Brauns, K. (2015). Computer versus paper--does it make any difference in test performance? *Teach. Learn. Med.* 27, 57–62. doi: 10.1080/10401334.2014.979175

Klassert, A. (2011). Lexikalische Fähigkeiten bilingualer Kinder mit Migrationshintergrund-Eine Studie zum Benennen von Nomen und Verben im Russischen und Deutschen [Lexical Abilities of Bilingual Children with an Immigrant Background – A Study of Naming of Nouns and Verbs in Russian and German]. Dissertation: University of Marburg.

Lee, K. S., Osborne, R. E., and Carpenter, D. N. (2010). Testing accommodations for university students with AD/HD: computerized vs. paper-pencil/regular vs. extended time. *J. Educ. Comput. Res.* 42, 443–458. doi: 10.2190/EC.42.4.e

Leeson, H. V. (2006). The mode effect: a literature review of human and technological issues in computerized testing. *Int. J. Test.* 6, 1–24. doi: 10.1207/s15327574ijt0601_1

Lenhard, A., Lenhard, W., and Schneider, W. (2020). ELFE II. Ein Leseverständnistest für Erst-bis Siebtklässler [A reading comprehension test for Grade 1 to 7 students]: Version II. Göttingen: Hogrefe, 4.

Lenhard, W., Schroeders, U., and Lenhard, A. (2017). Equivalence of screen versus print Reading comprehension depends on task complexity and proficiency. *Discourse Process.* 54, 427–445. doi: 10.1080/0163853X.2017.1319653

Melby-Lervåg, M., and Lervåg, A. (2014). Reading comprehension and its underlying components in second-language learners: a meta-analysis of studies comparing first-and second-language learners. *Psychol. Bull.* 140, 409–433. doi: 10.1037/a0033890

Merchant, G. (2015). Keep taking the tablets: iPads, story apps and early literacy. *AJLL* 38, 3–11. doi: 10.1007/BF03651950

Muter, V., Hulme, C., Snowling, M. J., and Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study. *Dev. Psychol.* 40, 665–681. doi: 10.1037/0012-1649.40.5.665

Muthén, L. K., and Muthén, B. O. (1998-2012). Mplus User's Guide. *7th* Edn. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., and Muthén, B. (1998-2017). Los Angeles, CA: Mplus User's Guide.

Neumann, M. M., Anthony, J. L., Erazo, N. A., and Neumann, D. L. (2019). Assessment and technology: mapping future directions in the early childhood classroom. *Front. Educ.* 4. doi: 10.3389/feduc.2019.00116

Paleczek, L., Seifert, S., and Schöffl, M. (2021). Comparing digital to print assessment of receptive vocabulary with GraWo-KiGa in Austrian kindergarten. *Br. J. Educ. Technol.* 52, 2145–2161. doi: 10.1111/bjet.13163

Perfetti, C. A., and Hart, L. (2002). The lexical quality hypothesis. *Precurs. Funct. Liter.* 11, 67–86. doi: 10.1075/swll.11.14per

Poggio, J., Glasnapp, D. R., Yang, X., and Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *J. Technol. Learn. Assesment* 3.

Puhan, P., Boughton, K., and Kim, S. (2007). Examining differences in examinee perfomance in paper and pencil and computerized testing. *J. Technol.* 6, 1–20.

Ricoy, M.-C., and Sánchez-Martínez, C. (2020). Revisión sistemática sobre el uso de la tableta en la etapa de educación primaria [A systematic review of tablet use in primary education]. *Revista Española Pedag.* 78, 273–290. doi: 10.22550/REP78-2-2020-04

Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*. 8, 23–74. doi: 10.23668/psycharchives.12784

Schöfl, M., Steinmair, G., Holzinger, D., and Weber, C. (2022). Predicting word reading deficits using an app-based screening tool at school entry. *Front. pediatr.* 10:863477. doi: 10.3389/fped.2022.863477

Seifert, S., and Paleczek, L. (2022). Comparing tablet and print mode of a German reading comprehension test in grade 3: Influence of test order, gender and language. *Int. J. Educ. Res.* 113. doi: 10.1016/j.ijer.2022.101948

Seifert, S., Paleczek, L., and Gasteiger-Klicpera, B. (2019). Rezeptive Wortschatzleistungen von Grundschulkindern mit Deutsch als Erst- und Zweitsprache und der Zusammenhang zu den Lesefähigkeiten: Implikationen für einen inklusiven Unterricht. *Empirische Sonderpädagogik*, 4, 259–278. doi: 10.25656/01:18334

Seifert, S., Paleczek, L., Schwab, S., and Gasteiger-Klicpera, B. (2017). *Grazer Wortschatztest – GraWo [Graz vocabulary test)*. Göttingen: Hogrefe.

Singer-Trakham, L. M., Alexander, P. A., and Berkowitz, L. E. (2019). Effects of processing time on comprehension and calibration in print and digital mediums. *J. Exp. Educ.* 87, 101–115. doi: 10.1080/00220973.2017.1411877

Statistik Austria. (2022). Bildung in Zahlen 2020/21- Schlüsselindikatoren und Analysen. Wien: Statistik Austria.

Steedle, J. T., Cho, Y. W., Wang, S., Arthur, A. M., and Li, D. (2022). Mode effects in college admissions testing and differential Speededness as a possible explanation. *Educ. Meas. Issues Pract.* 41, 14–25. doi: 10.1111/emip.12484

Støle, H., Mangen, A., and Schwippert, K. (2020). Assessing children's reading comprehension on paper and screen: a mode-effect study. *Comput. Educ.* 151:103861. doi: 10.1016/j.compedu.2020.103861

Taherbhai, H., Seo, D., and Bowman, T. (2012). Comparison of paper-pencil and online performances of students with learning disabilities. *Br. Educ. Res. J.* 38, 61–74. doi: 10.1080/01411926.2010.526193

Trautwein, J., and Schroeder, S. (2019). WOR-TE: Ein Ja/Nein-Wortschatztest für Kinder verschiedener Altersgruppen. *Diagnostica* 65, 37–48. doi: 10.1026/0012-1924/a000212

Wagner, I., Loesche, P., and Bißantz, S. (2022). Low-stakes performance testing in Germany by the VERA assessment: analysis of the mode effects between computer-based testing and paper-pencil testing. *Eur. J. Psychol. Educ.* 37, 531–549. doi: 10.1007/s10212-021-00532-6

Wang, S., Jiao, H., Young, M. J., Brooks, T., and Olson, J. (2007). A Meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educ. Psychol. Meas.* 67, 219–238. doi: 10.1177/0013164406288166

Wang, S., Jiao, H., Young, M. J., Brooks, T., and Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 Reading assessments. *Educ. Psychol. Meas.* 68, 5–24. doi: 10.1177/0013164407305592

Wang, T.-H., Kao, C.-H., and Chen, H.-C. (2021). Factors associated with the equivalence of the scores of computer-based test and paper-and-pencil test: presentation type, item difficulty and administration order. *Sustain. For.* 13:9548. doi: 10.3390/su13179548

Wendt, H., and Schwippert, K. (2017). "Lesekompetenzen von Schülerinnen und Schülern mit und ohne Migrationshintergrund" in IGLU 2016: Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. eds. A. Hußmann, H. Wendt, W. Bos, A. Bremerich-Vos, D. Kasper and E.-M. Lankeset al. (Münster, New York: Waxmann), 219–234.

Yu, C.-Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. Dissertation. University of California, Los Angeles, CA.