Original Article

Effects of Test Mode and Medium on Elementary School Students' Test Experience

Thomas Brüggemann, Ulrich Ludewig, Ramona Lorenz, and Nele McElvany

Center for Research on Education and School Development, TU Dortmund University, Dortmund, Germany

Abstract. The use of digital media in education can bring great benefits and its use in schooling is steadily increasing. Administrating paper-versus computer-based as well as fixed-item versus adaptive tests could create differences in test experience, which can threaten the comparability of test results. This study investigated how the pen-and-paper, computer-based, and computer-adaptive test formats of a standardized reading comprehension test affect test anxiety and motivation among German fourth-grade students. A within-class randomized field trial with 387 fourth graders (aged: 9–10 years; 46.3% female) was conducted. Repeated-measures analysis of covariance (ANCOVA) revealed no differences in state test anxiety between the test formats when controlling for trait test anxiety and pre-test state anxiety, but state reading motivation was initially higher when reading on a screen, controlling for trait reading motivation. However, this difference diminishes over the course of the test. Implications for using digital media in elementary school test situations are discussed.

Keywords: primary education, anxiety, motivation, mode effect, adaptive testing

Digital media offer opportunities for teachers to improve the teaching, monitoring, and evaluation of their student's learning progress. Digitalization has particularly advanced the field of student assessment. Replacing pen-and-paper tests (PPT), digitally administered computer-based tests (CBT) and computer adaptive tests (CAT) have recently increased in popularity. For instance, large-scale assessments (LSA) like the Progress of International Reading Literacy Study (PIRLS) are switching to CBTs (Hußmann et al., 2017), while the Programme for International Student Assessment (PISA) is adopting adaptive tests (Yamamoto et al, 2019). CATs estimate an examinee's ability throughout the test and use this estimate to administer items that provide the highest amount of information to estimate their ability, making them more efficient than fixed-item tests (FIT; e.g., Ling et al., 2017). However, doubts about the direct comparability of these formats with regard to the test experience have been raised (e.g., Chua, 2012; Colwell, 2013).

Most research on test experience compares only two of the three test formats (PPT, CBT, or CAT) simultaneously. For instance, Chua (2012) compared the test motivation experienced in PPTs and CBTs and Martin and Lazendic (2018) investigated test anxiety in CBTs and CATs. In contrast, the effects of test formats on the test experience of elementary school students taking tests of reading comprehension, which is an important predictor for educational success (e.g., Schwabe et al., 2015), are fairly fairly unexplored. This study presents a systematic experimental study investigating the effects of different test formats on test experience among elementary school students during and after a reading comprehension test.

Test Experience

When confronted with a test within a given time frame in a classroom setting, students can experience test anxiety (von der Embse et al., 2018) as well as test motivation (Chua, 2012; Weiss & Betz, 1973). Test anxiety refers to a set of emotional, physiological, and behavioral responses that accompany a person's concerns about possible negative consequences of failure on a test or other evaluative situation (Sieber et al., 1977). Test motivation or test-taking motivation is a specific form of achievement motivation and can be conceptualized as a situation-specific motivation to perform well in an evaluative situation (Baumert & Demmrich, 2001). During a test, both test anxiety and test motivation are tied to the specific situation (state) and are dependent on people's predispositions (traits). While states are

European Journal of Psychological Assessment (2023)
https://doi.org/10.1027/1015-5759/a000767

situation-specific and bound to a particular point in time, traits are stable attributes (Tremblay et al., 1995). When investigating the effects of test formats on states, it is important to consider the role of the associated traits (e.g., Tremblay et al., 1995; Zohar, 1998).

Test Anxiety

In the additive model of test anxiety, test anxiety is understood as consisting of state and trait test anxiety (Zohar, 1998). State test anxiety can be influenced by several factors, such as the perceived importance of the test, the examinee's preparedness, or the level of self-confidence. An individual's state test anxiety results from their trait test anxiety as well as situation-specific variables. Higher levels of trait test anxiety lead to higher levels of state test anxiety (Paulman & Kennelly, 1984).

Differences in Test Anxiety Between Test Media

Older studies reported that CBT can elicit test anxiety, often in conjunction with computer anxiety (e.g., Shermis & Lombard, 1998). However, the now-ubiquitous presence of computers has made computer anxiety less relevant for children today (dos Santos & de Santana, 2018). Recently, Sahlan et al. (2021) found no differences in test anxiety between PPTs and CBTs for 11- to 16-year-old high school students taking an English language test.

Differences in Test Anxiety Between Test Modes

Adaptive test administration could influence text anxiety (Colwell, 2013). In a CAT, all examinees answer approximately the same proportion of items correctly, regardless of their ability. However, examinees often expect to answer more or fewer items correctly based on their habitual expectations. This disconnect between the proportion of items answered correctly and the examinees' expectations can negatively affect an examinee's test anxiety and motivation (Tonidandel et al., 2002). Ling et al. (2017) compared the effects of different configurations of CATs and FITs in a short mathematics test for grades six to eight. Participants reported lower levels of state test anxiety in the FIT condition than in the CAT. Similarly, Martin and Lazendic (2018) found that 3rd to 9th-grade students experienced higher levels of state test anxiety in the CAT condition than in the CBT condition for a mathematics test. Ortner and Caspers (2011) discovered an interaction effect between trait test anxiety and administration mode. They observed that examinees with high levels of trait test anxiety performed worse in a CAT than in a FIT and concluded that adaptive tests may exhibit biases to the disadvantage of examinees with high levels of test anxiety.

In conclusion, there is reason to believe that test formats may differ in terms of how much test anxiety examinees experience. However, the effects of different test formats on experienced state anxiety among young readers are inconclusive.

Test Motivation

Similar to anxiety, state motivation is a function of trait motivation and situational characteristics (Tremblay et al., 1995). This view extends to test-taking motivation (Helm & Warwas, 2018). In the case of a reading comprehension test, state reading motivation is highly relevant. As a motivational state, it is tied to a reading task and assesses a student's willingness to engage with the test subject in the form of reading test items. As such, a student's state of reading motivation within a reading comprehension test reflects their motivation to engage with the test items and hence with the test itself (Lepper et al., 2021). State reading motivation is affected by situational characteristics and individual traits.

Differences in Test Motivation Between Test Media

There are reasons to expect that testing on a screen can improve motivation among elementary school students. Empirical studies have found that digital media can increase children's motivation to read (Picton, 2014) as well as their test-taking motivation (Chua, 2012). One probable cause is the *novelty effect* (Shin et al., 2019), which states that new experiences can generate positive attitudes simply due to their novelty. Though today's students are more accustomed to digital media, schools in Germany rarely make use of computers for testing purposes (Fraillon et al., 2020).

Differences in Test Motivation Between Test Modes

In the 1970s, Weiss and Betz (1973) argued that CATs could have beneficial effects on test motivation, as CATs challenge high-ability students more and discourage low-ability students less, thus increasing motivation. However, evidence for this view has been lacking. Both Frey et al. (2009) in a concentration test and Ortner et al. (2014) in a reasoning test found small negative effects of CATs on test motivation measured as the perceived probability of success. For elementary school students, Martin and Lazendic (2018) found no significant difference between test modes on motivation in a mathematics test. Nevertheless, there are plausible arguments for the effects of CATs on test motivation, although the exact nature of such effects is unclear, especially with regard to reading comprehension tests.

Current Study

Previous research on how test formats can affect the test experience of young readers taking a reading comprehension test is limited. It is important to investigate whether theoretical expectations stemming from studies with older students hold for young students who have less experience with tests. Hence, this study investigates the effects of the PPT, CBT, and CAT test formats on the test experience of 4th-grade students via the following research questions:

> *Research Question 1 (RQ1)*: To what extent does administering a reading comprehension test as a PPT, CBT, or CAT affect the test anxiety of fourth grade students?
>
> *Research Question 2 (RQ2)*: To what extent does administering a reading comprehension test as a PPT, CBT, or CAT affect the reading motivation of fourth grade students?

Based on the argument by Colwell (2013) that CATs may increase test anxiety, the first hypothesis states that fourth-grade students' state test anxiety will be higher in a CAT than in a PPT or CBT, with no differences between the PPT and CBT (Hypothesis 1, H1). The second hypothesis states that reading motivation is lower in the PPT condition than in the CBT and CAT (Hypothesis 2a, H2a), based on the assumption that digital test formats increase students' motivation (Chua, 2012). Due to the *novelty effect*, it is furthermore expected that this difference is greater in the middle of the test than at its end (Hypothesis 2b, H2b).

Method

Participants

To investigate the hypotheses, 526 German fourth-grade students from 27 classes in 12 different elementary schools in western Germany were sampled between October 2020 and December 2021. The operational sample consisted of N = 387 students (46.3% female) who provided parental consent. The data from the 139 students who did not have parental consent was deleted. University ethics approval was obtained (GEKTEDO_2020_26). The students' mean age was 9.53 (SD = 0.66) years; 13.5% were not born in Germany. Students were assigned to one of three experimental groups within their class at random (N_{PPT} = 120, N_{CBT} = 135, N_{CAT} = 132). There were no differences between the students in the three test formats regarding gender, $\chi^2(2)$ = 3.35, p = .187, country of birth (native-born or immigrant), $\chi^2(2)$ = 4.67, p = .097, test performance, F(2, 384) = 0.05, p = .950, or mid-year grade in German language arts, F(2, 272) = 1.73, p = .179, as an indicator of students' academic performance level.

Instruments

Test Anxiety

Trait test anxiety was measured with a shortened version of the German Test Anxiety Inventory (TAI-G; Wacker et al., 2008) by Bertrams and Englert (2014). There are five items for worry and four for emotionality. State anxiety was measured with the State-Trait Anxiety Inventory State-Kurzskala-Deutsch (STAI-SKD), which consists of two items for worry and three for emotionality, though differential analyses of the two subscales are not recommended (Englert et al., 2011).

Reading Motivation

Trait reading motivation was measured with three statements from the reading motivation scales used in PIRLS 2016 (Hußmann et al., 2017). A fourth item ("I enjoy reading") was added to the scale. The state reading motivation scale was based on a four-item scale used in the German national supplementary test for PISA 2000 (Kunter et al., 2002) and adapted to the reading task at hand (Lepper et al., 2021).

Reading Comprehension

The Faire und adaptive Lesekompetenzdiagnose (FALKE) is a reading comprehension test for German third- and fourth-grade students (Ludewig et al., 2021). It consists of 44 narratives and 41 expository texts with a mean length of 60.15 (SD = 17.67) words. There are 69 text-based and 63 inference-based calibrated multiple-choice items. Some items share a text, though texts are not administered twice. In all conditions, 25 items were administered. For the FIT (i.e. both the PPT and CBT), items were selected based on simulations of the adaptive version of the test using the R-library catR (Magis & Raîche, 2012) in R 4.1.2 (R Core Team, 2022). Selection criteria for items were the probability of being chosen for the adaptive test and difficulty, text type, and question type. FIT item difficulties were normally distributed.

The test formats were held visually and conceptually equivalent, with the PPT printed on ISO 216 A5 sheets in landscape format, one item per page. The constraints of the CAT were also implemented for the CBT and PPT. Specifically, students were unable to return to a previous item in both the CBT and PPT versions, as this was not possible in the CAT.

To account for rapid guessing behavior, responses in the CBT and CAT within four seconds were considered as not administered (see Wise & DeMars, 2006). Reading comprehension scores were calculated with weighted likelihood estimates (Warm, 1989) with fixed calibrated item parameters using the library catR (Magis & Raîche, 2012). The WLE reliability of the reading comprehension test was good for all test formats (PPT: Rel._{WLE} = .82; CBT: Rel._{WLE} = .80;

CAT: Rel.$_{WLE}$ = .89). More detailed descriptions of the measures can be found in the supplementary materials (ESM 1, Brüggemann, 2023).

Procedure

The study used both a within-subject and a between-subject design. Students were tested in their classrooms by two trained test administrators. Participants were assigned to one of the three formats (PPT, CBT, and CAT) at random. Prepared at each student's desk were a tablet and a paper-based introductory questionnaire containing trait scales for anxiety and reading motivation as well as questions on sociodemographic variables. After all students had finished the introductory pre-test questionnaire, participants were informed that they would be completing a reading comprehension test. Immediately afterwards, they were asked to rate their pre-test state anxiety. Only afterward did students receive their assigned test format and commence with the reading comprehension test. After the 12th item (midway), as well as after the last item (post-test), the state measures for anxiety and reading motivation were administered.

Analyses

In order to investigate differences between the test formats regarding test anxiety (H1), a repeated-measures analysis of covariance (ANCOVA) with midway and post-test state anxiety as repeated measures, test format as a between-subjects factor and trait anxiety and pre-test state anxiety as covariates were calculated. A repeated-measures ANCOVA with state reading motivation as a within-subject factor, test format as a between-subjects factor, and trait reading motivation as a covariate was used to test H2a and H2b. Analyses were conducted with IBM SPSS 28.

Due to low rates of missing values, listwise deletion was used. Missing values were lowest for the midway-state anxiety measure (2.3%) and highest for the post-test reading motivation scale (6.5%).

Results

Manipulation Check

We compared the mean of the average percentage of correct answers of the students in the FITs and the CAT in order to check whether the CAT performed as expected. The variance of the percentage of correct answers per student should be substantially lower in the CAT than in the FITs. In FITs, students' percentage of correct answers varies as a function of their ability. In a perfectly adaptive CAT, all students should have the same percentage of correct answers because the CAT adapts the item to the students' ability. However, in practice, the item pool is limited; thus, students can receive items that are not perfectly matching their estimated ability. Therefore, the difference in variance in percentage of correct answers between FIT and CAT indicates the degree of the adaptivity of the CAT. The variance in the percentages of correct answers was statistically significantly lower in the CAT (σ^2 = 3.82) than in the FITS (σ^2 = 5.21), F(1, 383) = 6.19, p = .013, indicating that the CAT performed as expected. The percentage of correct answers in the FITs (M = 64.5%, SD = 2.28) and the CAT (M = 60.6%, SD = 1.96) per student did not differ statistically significantly, t(303) = 1.76, p = .079.

Test Experience

Descriptive statistics for the different measures of test experience are displayed in Table 1. The scales were generally reliable (α_{min} = .67, α_{max} = .91). Pre-test anxiety was descriptively highest in the adaptive condition, but the differences were not statistically significant, F(2, 369) = 2.78, p = .063. There was no difference between the test formats regarding trait test anxiety, F(372, 2) = 0.42, p = .651, or trait reading motivation, F(372, 2) = 0.69, p = .650.

Correlations between corresponding trait and state measures were weak to moderate in the expected directions, with trait anxiety correlating weakly with the state anxiety measures (r_{min} = .356), and the state anxiety measures correlating moderately with each other (r_{min} = .539). Similarly, the correlation between trait and state reading motivation was weak (r_{min} = .329), while midway and post-test reading motivation correlated strongly with each other (r = .849). Intercorrelations among all variables can be found in the supplementary materials (Table S2 in ESM 2, Brüggemann, 2023).

Test Anxiety

Repeated measures ANCOVA found no effects of test format on state anxiety. There was no statistically significant difference between the test formats, F(2, 345) = 1.88, p = .155; test anxiety did not differ between measurement points, F(1, 345) = .02, p = .896; nor was there an interaction effect between state anxiety and test format, F(2, 345) = 0.07, p = .935. State anxiety was predicted by both covariates for pre-test anxiety, F(2.345) = 73.14, p < .001, and trait anxiety, F(2, 345) = 16.94, p < .001. Figure 1 shows the estimated marginal means of state anxiety over the course of the test for all three test formats. There were no statistically significant differences between the test formats regarding test anxiety, leading us to reject H1.

Reading Motivation

The second repeated-measures ANCOVA found that state reading motivation did not differ between the measurement

Table 1. Means (*M*), standard deviations (*SD*), and reliability coefficients (α) for the test experience measures of trait test anxiety, state test anxiety, trait reading motivation, and state reading motivation by test format and measurement point

		Test anxiety			Reading motivation		
Test format	Measurement point	*M*	*SD*	α	*M*	*SD*	α
PPT	Trait	2.26	0.66	.83	3.27	0.87	.88
	State pre	1.96	0.68	.69			
	State midway	1.83	0.69	.76	2.66	0.87	.88
	State post	1.90	0.84	.85	2.71	0.96	.91
CBT	Trait	2.25	0.65	.83	3.35	0.84	.91
	State pre	1.98	0.66	.67			
	State midway	1.96	0.68	.67	2.92	0.78	.79
	State post	2.00	0.82	.78	2.79	0.93	.89
CAT	Trait	2.32	0.64	.82	3.26	0.82	.86
	State pre	2.14	0.68	.67			
	State midway	2.05	0.69	.68	3.03	0.80	.78
	State post	2.13	0.90	.82	2.90	0.96	.90
Total	Trait	2.27	0.65	.83	3.29	0.83	.89
	State pre	2.03	0.68	.68			
	State midway	1.95	0.69	.71	2.88	0.82	.82
	State post	2.02	0.86	.82	2.80	0.95	.90

Note. PPT = pen and paper test; CBT = computer based test; CAT = computer adaptive test. N_{total} = 387, N_{PPT} = 120, N_{CBT} = 135, N_{CAT} = 132; sample sizes for the descriptive statistics range from N_{min} = 113 to N_{max} = 378; state reading motivation was not measured at the pre-test measurement point.

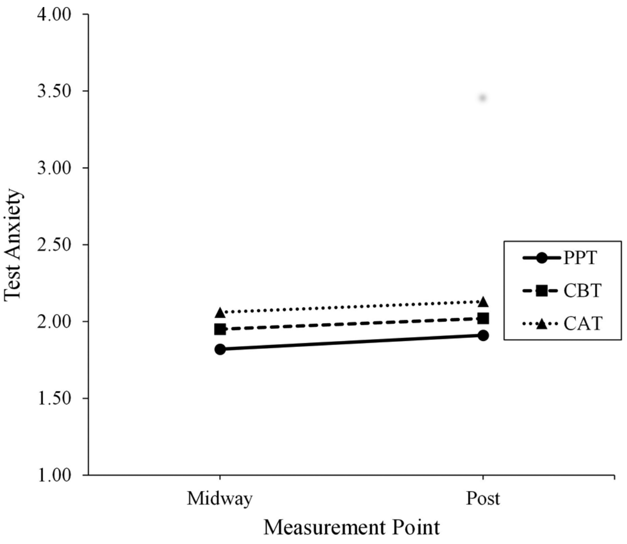

Figure 1. State test anxiety in the test formats over the measurement points. This figure shows the estimated marginal means of state test anxiety for the test formats over the measurement points, controlling for pre-test anxiety and trait anxiety. PPT = pen-and-paper test; CBT = computer-based test; CAT = computer adaptive test.

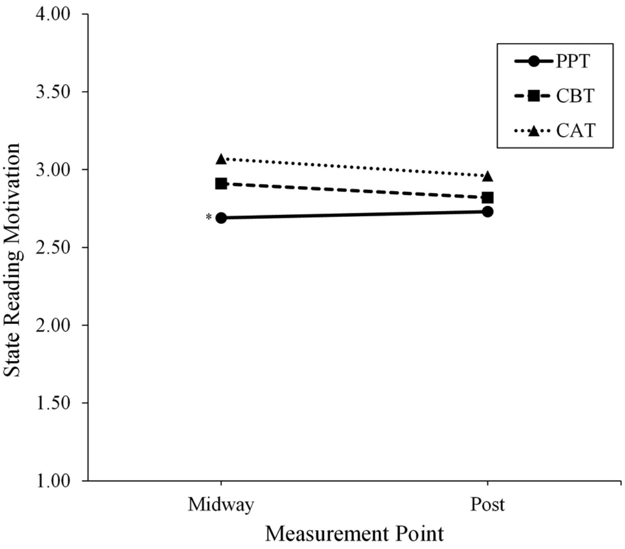

Figure 2. State reading motivation in the test formats over the measurement points. This figure shows the estimated marginal means for state reading motivation for the test formats over the measurement points, controlling for trait reading motivation. PPT = pen-and-paper test; CBT = computer-based test; CAT = computer adaptive test. *Midway state reading motivation was significantly lower in the PPT than in the CAT.

points, F(1, 339) = 2.88, p = .091, but a statistically significant interaction between state reading motivation and test format indicated that state reading motivation developed differently over time between the test formats, F(2, 339) = 3.20, p = .042; η_p^2 = .019. In addition, there was a statistically significant difference between the test formats in state reading motivation, F(2, 339) = 4.02, p = .019; η_p^2 = .023. Trait reading motivation was a significant predictor for state reading motivation, F(1, 339) = 45.27, p < .001. Figure 2 shows the development of state reading motivation over the course of the test. It can be seen that state reading motivation was higher in the computer-based test formats, supporting H1a and that motivation diminished over the course of the test in the CBT and CAT conditions, but not in the PPT condition, which is in line with H2b.

Discussion

In this study, the effects of different test formats, namely paper-based (PPT), computer-based (CBT), and computer adaptive testing (CAT), on the test experience of 387 fourth-grade students taking a reading comprehension test were investigated in a quasi-experimental within- and between-subject design. The results showed no differences in state test anxiety between the test formats. State reading motivation was initially higher when the test was administered on a screen (i.e. CBT or CAT), although the differences subsided over the course of the test.

Test Anxiety

The analyses regarding test anxiety found no statistically significant differences between the test formats or differential effects of the test formats with regard to the development of state test anxiety over the course of the test. Trait anxiety was a significant predictor for state test anxiety, which conforms to the additive model by Zohar (1998). Hypothesis 1, which stated that test anxiety would be higher in the CAT than in the FITs, had to be rejected. This result is surprising considering that previous research found that students in the adaptive testing condition experienced higher levels of test anxiety (Ling et al., 2017). Fourth-grade students may be less sensitive to the administration differences between a FIT and a CAT than older students. They may be less experienced with tests and have different expectations than older students, who might have stronger habitual expectations regarding their own performance and established preferences regarding test features, as described by Colwell (2013).

Reading Motivation

Reading motivation was investigated in Hypothesis H2, which assumed higher levels of reading motivation among students tested on a computer rather than on paper. The results showed statistically significantly lower levels of state reading motivation for students in the PPT at the midway point of the test, affirming H2a, although the statistically significant interaction effect indicated that this effect diminished over the course of the test, supporting H2b. This finding conforms to previous research on the motivating effects of digital media (Chua, 2012). There were no differences in test motivation between CBTs and CATs, which is in line with recent research (Martin & Lazendic, 2018). The *novelty effect* suggests that experiencing new stimuli leads to a positive affect simply because the stimuli are new (Shin et al., 2019). Using computers in the classroom for testing purposes may have been a new experience for the students at first, initially increasing their reading motivation. However, over the course of the test, the students got used to the computers, causing a decline in motivation to a similar level as the students taking the PPT.

Strengths and Limitations

Data collection for this study was undertaken during the height of the COVID-19 pandemic, which forced schools to close intermittently. Thus, the tested students may have been more confident and experienced in using computers as a learning tool than fourth-grade students in the past. Though this may affect the study's comparability with previous studies, it makes the results more relevant for a future in which students are more experienced with digital media for learning. The test environments were low-stakes, which might affect the results generalizability to high-stakes test situations, but does make them relevant for large-scale assessments. Furthermore, this study did not consider differential effects of test performance on the test experience (for a brief discussion, see supplementary materials [ESM 2], Brüggemann, 2023). Lastly, a strength of the study was the unique experimental design comparing three test formats in parallel within a class, with within-subject measures to investigate the development of the dependent variables within individual students over time. Additionally, the manipulation check confirmed the efficacy of the administered CAT, and the within-class design allowed for comparisons even though data collection was stretched over two school years.

Conclusion

There is much we do not know about how test administration affects the test experience. This study concludes that test equivalence between PPTs, CBTs, and CATs in terms of test experience is achievable for young readers. CBTs and CATs do not seem to increase these students' test anxiety relative to PPTs. Instead, students are initially more motivated when being tested on a computer, though the effect wanes over time. Hence, digital media in education can yield a temporary increase in students' motivation, which can be used for instructional purposes. More efficient test formats, such as CATs, can further limit the reduction in motivation over time by allowing for shorter tests. Therefore, the results of this study should encourage more widespread use of computer-based and computer-adaptive tests for reading comprehension assessment in elementary schools in low-stakes situations. Future research could look into the effects in high-stakes situations, the longevity of the motivational increase in the case of repeated computer-based test administration, and the effects of age on the test experience.

References

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441–462. https://doi.org/10.1007/BF03173192

Bertrams, A., & Englert, C. (2014). Test anxiety, self-control, and knowledge retrieval in secondary school students. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie, 46*(4), 165–170. https://doi.org/10.1026/0049-8637/a000111

Brüggemann, T. (2023). *Supplement to "Effects of test mode and medium on elementary school students' test experience"*. https://osf.io/76hc2/?view_only=baf5985b8cb94fe4b78cb-ce4261aee7a. https://doi.org/10.17605/OSF.IO/76HC2

Chua, Y. P. (2012). Effects of computer-based testing on test performance and testing motivation. *Computers in Human Behavior, 28*(5), 1580–1586. https://doi.org/10.1016/j.chb.2012.03.020

Colwell, N. M. (2013). Test anxiety, computer-adaptive testing, and the common core. *Journal of Education and Training Studies, 1*(2), 50–60. https://doi.org/10.11114/jets.vli2.101

dos Santos, T. D., & de Santana, V. F. (2018). Computer anxiety and interaction: A systematic review. In *Proceedings of the 15th International Web for All Conference* (pp. 1–10). https://doi.org/10.1145/3192714.3192825

Englert, C., Bertrams, A., & Dickhäuser, O. (2011). Entwicklung der Fünf-Item Kurzskala STAI-SKD zur Messung von Zustandsangst [Development of the five-item short scale STAI-SKD for the assessment of state anxiety]. *Zeitschrift für Gesundheitspsychologie, 19*, 173–180. https://doi.org/10.1026/0943-8149/a000049

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020). *Preparing for life in a digital world: IEA international computer and information literacy study 2018 international report*. Springer Nature. https://doi.org/10.1007/978-3-030-38781-5

Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests [Effects of adaptive testing on test-taking motivation with the example of the Frankfurt Adaptive Concentration Test]. *Diagnostica, 55*(1), 20–28. https://doi.org/10.1026/0012-1924.55.1.20

Helm, C., & Warwas, J. (2018). Psychological determinants of test motivation in low-stakes test situations: A longitudinal study of singletrait–multistate models in accounting. *Empirical Research in Vocational Education and Training, 10*(1), 1–34. https://doi.org/10.1186/s40461-018-0074-7

Hußmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E.-M., McElvany, N., Stubbe, T. C., & Valtin, R. (2017). *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* [International comparison of German elementary school students' reading competencies]. Waxmann.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillman, K.-J., & Weiß, M. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente* (Materialien aus der Bildungsforschung Nr. 72) [PISA 2000: Documentation of the survey instruments (Educational research materials No. 72)]. Max-Planck-Institut für Bildungsforschung.

Lepper, C., Stang, J., & McElvany, N. (2021). Gender differences in text-based interest: Text characteristics as underlying variables. *Reading Research Quarterly, 57*(2), 537–554. https://doi.org/10.1002/rrq.420

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement, 41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Ludewig, U., Trendtel, M., Schlitter, T., & McElvany, N. (2021). Adaptives Testen von Textverständnis in der Grundschule [Adaptive Testing of Text Comprehension in Primary School. Development of a CAT-Optimized Item Pool]. *Diagnostica, 68*(1), 39–50.

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software, 48*(8), 1–31. https://doi.org/10.18637/jss.v048.i08

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology, 110*(1), 27–45. https://doi.org/10.1037/edu0000205

Ortner, T. M., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment, 27*(3), 157–163. https://doi.org/10.1027/1015-5759/a000062

Ortner, T. M., Weißkopf, E., & Koch, T. (2014). I will probably fail. *European Journal of Psychological Assessment, 30*(1), 48–56. https://doi.org/10.1027/1015-5759/a000168

Paulman, R. G., & Kennelly, K. J. (1984). Test anxiety and ineffective test taking: Different names, same construct? *Journal of Educational Psychology, 76*(2), 279–288. https://doi.org/10.1037/0022-0663.76.2.279

Picton, I. (2014). *The impact of ebooks on the reading motivation and reading skills of children and young people: A rapid literature review*. National Literacy Trust. Retrieved October 13, 2022 from https://eric.ed.gov/?id=ed560635

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Sahlan, F., Alberth., Madil, W., & Hutnisyawati. (2021). The effects of modes of test administration on test anxiety and test scores: A study in an Indonesian school. *Issues in Educational Research, 31*(3), 952–971. https://www.iier.org.au/iier31/sahlan.pdf

Schwabe, F., McElvany, N., & Trendtel, M. (2015). The school age gender gap in reading achievement: Examining the influences of item format and intrinsic reading motivation. *Reading Research Quarterly, 50*(2), 219–232. https://doi.org/10.1002/rrq.92

Shermis, M. D., & Lombard, D. (1998). Effects of computer-based test administrations on test anxiety and performance. *Computers in Human Behavior, 14*(1), 111–123. https://doi.org/10.1016/S0747-5632(97)00035-6

Shin, G., Feng, Y., Jarrahi, M. H., & Gafinowitz, N. (2019). Beyond novelty effect: A mixed-methods exploration into the motivation for long-term activity tracker use. *JAMIA Open, 2*(1), 62–72. https://doi.org/10.1093/jamiaopen/ooy048

Sieber, J. E., O'Neil, J., & Tobias, S. (1977). *Anxiety, learning, and instruction*. Routledge. https://doi.org/10.4324/9780203056684

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87*(2), 320–332. https://doi.org/10.1037/0021-9010.87.2.320

Tremblay, P. F., Goldberg, M. P., & Gardner, R. C. (1995). Trait and state motivation and the acquisition of Hebrew vocabulary. *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement, 27*(3), 356–370. https://doi.org/10.1037/0008-400X.27.3.356

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders, 227*, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

Wacker, A., Jaunzeme, J., & Jaksztat, S. (2008). Eine Kurzform des Prüfungsängstlichkeitsinventars TAI-G [A short version of the Test Anxiety Inventory TAI-G]. *Zeitschrift für Pädagogische Psychologie, 22*(1), 73–81. https://doi.org/10.1024/1010-0652.22.1.73

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/bf02294627

Weiss, D. J., & Betz, N. E. (1973). *Ability measurement: Conventional or adaptive?* (Research Report 73-1). Department of Psychology, University of Minnesota.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Yamamoto, K., Shin, H., & Khorramdel, L. (2019). *Introduction of multistage adaptive testing design in PISA 2018* (OECD Education Working Papers No. 209). OECD Publishing. https://doi.org/10.1787/b9435d4b-en

Zohar, D. (1998). An additive model of test anxiety: Role of exam-specific expectations. *Journal of Educational Psychology, 90*(2), 330–340. https://doi.org/10.1037/0022-0663.90.2.330

History

Received June 27, 2022
Revision received February 1, 2023
Accepted February 16, 2023
Published online May 23, 2023
EJPA Section / Category Educational Psychology

Conflict of Interest

We have no conflicts of interest to disclose.

Publication Ethics

The approval of the ethics committee of the department of Educational Sciences and Psychology at the TU Dortmund University was obtained (GEKTEDO_2020_26).

Authorship

Thomas Brüggemann, Conceptualization, Methodology, Investigation, Software, Writing – Original Draft; Ulrich Ludewig, Conceptualization, Methodology, Supervision, Writing – Review & Editing; Ramona Lorenz, Conceptualization, Supervision, Writing – Review & Editing; Nele McElvany, Writing – Review & Editing, Supervision.

Open Science

The supplementary materials can be found at https://osf.io/76hc2/?view_only=baf5985b8cb94fe4b78cbce4261aee7a (Brüggemann, 2023). We report how we determined our sample size, all data exclusions, all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures used 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: I confirm that there is sufficient information for an independent researcher to reproduce all of the reported results, including codebook. The data can be found in the supplementary materials. Available is an SPSS data set, a codebook, and a corresponding SPSS syntax file to replicate the analyses presented in the paper (Brüggemann, 2023).

Open Materials: The information needed to reproduce all of the reported methodology is not openly accessible. The item and scale descriptions can be found in the supplementary materials (Brüggemann, 2023). The reading comprehension test is available on request from the authors.

Preregistration of Studies and Analysis Plans: This study was not preregistered with an analysis plan.

Funding

This work was supported by the German Federal Ministry of Education and Research under Grant Number 01JD1813. Open access publication enabled by TU Dortmund University.

ORCID

Thomas U. Brüggemann
https://orcid.org/0000-0003-1381-5025
Ramona Lorenz
https://orcid.org/0000-0002-5733-5421
Ulrich Ludewig
https://orcid.org/0000-0001-9614-847X

Thomas Brüggemann
Center for Research on Education and School Development
TU Dortmund University
Vogelpothsweg 78
44227 Dortmund
Germany
thomas.brueggemann@udo.edu