European Journal of Psychology of Education (2022) 37:531–549
https://doi.org/10.1007/s10212-021-00532-6

Low-stakes performance testing in Germany by the VERA assessment: analysis of the mode effects between computer-based testing and paper-pencil testing

Inga Wagner[1] · Philipp Loesche[1] · Steven Bißantz[1]

Received: 28 August 2020 / Revised: 5 January 2021 / Accepted: 6 January 2021 /
Published online: 29 March 2021

Abstract
The German school system employs centrally organized performance assessments (some of which are called "VERA") as a way of promoting lesson development. In recent years, several German federal states introduced a computer-based performance testing system which will replace the paper-pencil testing system in the future. Scores from computer-based testing are required to be equivalent to paper-pencil testing scores so that the new testing medium does not lead to disadvantages for students. Therefore, the current study aimed at investigating the size of the mode effect and the moderating impact of students' gender, academic achievement and mainly spoken language in everyday life. In addition, the variance of the mode effect across tasks was investigated. The study was conducted in four German federal states in 2019 using a field experimental design. The test scores of 5140 eighth-graders from 165 schools in the subject German were analysed. The results of multi-level modelling revealed that students' test scores in the computerized version of the VERA test were significantly lower than in the paper-pencil version. Students with a lower academic achievement were more disadvantaged by the VERA computerized test. The results were inconsistent regarding the interactions between testing mode and students' gender and mainly spoken language in everyday life. The variance of the mode effect across tasks was high. Research into different subjects and in other federal states and countries under different testing conditions might yield further evidence about the generalizability of these results.

Keywords Computer-based testing · Paper-pencil testing · Performance tests · Lesson development

✉ Inga Wagner
iwagner@zepf.uni–landau.de

Extended author information available on the last page of the article

Introduction

Performance tests were introduced in Germany about 15 years ago as part of an external evaluation system conducted in schools. These tests are called "VERA" in several German federal states. They are implemented in Grade 3 in the subjects Mathematics and German and in Grade 8 in the subjects Mathematics, German, English and French (Kultusministerkonferenz, 2016).

Policymakers introduced the VERA tests in order to support school development and improvement of classroom practices (Kultusministerkonferenz, 2016). By reflecting on the VERA test results, teachers are supposed to obtain a better insight into their classes' deficits. Based on this knowledge they can adapt their teaching, e.g. by repeating content and tasks that were poorly solved by their students in the VERA tests. In contrary to the United States, where poor results in performance tests can be sanctioned (e.g. Stecher, 2002), there are no severe consequences for teachers of low-performing classes in Germany.

In recent years, several German federal states decided to convert the testing mode of the VERA assessment from paper-pencil testing to computer-based testing. Computer-based testing reduces copy costs, increases testing efficiency and generates immediate scores for multiple-choice items (Ejim, 2017). If the practical preparation and execution of VERA tests was less laborious, teachers' acceptance of the tests might increase. They would have more time and might be more motivated to engage in conceptually relevant activities such as reflecting on their students' test results. According to the sequential model for lesson development proposed by Helmke (2012), a better acceptance of performance tests and a deeper reflection of test results are positive prerequisites for favourable changes in teaching practices.

However, the introduction of computer-based testing requires that the scores are comparable with those obtained from paper-pencil testing so that students are not disadvantaged by this new testing medium. If this were not the case and if there was evidence that paper-based tests and computer-based tests did not produce the same results, there would be a mode effect between the two testing mediums (Clariana & Wallace, 2002). This issue is especially important when the results of performance tests are to be compared over time, or when computer-based performance testing is only partly introduced, as is the case in Germany, and comparisons between groups tested with different modes should still be possible (Bennett et al., 2008).

Therefore, the current study investigates whether scores from VERA computer-based testing differ from scores from paper-pencil testing. Student characteristics, such as gender, academic achievement and mainly spoken language in everyday life, which might moderate the size of the mode effect, are also taken into consideration. The variation of the mode effect across tasks is also examined. The study was conducted using a field experimental design in four German federal states.

Theoretical background

Advantages and disadvantages of computer-based testing

One practical advantage of computer-based testing is that printing and storing test booklets is no longer necessary (Thompson & Weiss, 2009). With computer-based testing, constructs can be measured with a higher validity and objectivity: Having the instructions and tasks read to

participants by a computer system, as is necessary with preschool children, leads to a higher degree of standardization than having the materials read by teachers. Automated scoring of test answers by a computer system is also more objective than scores made by different teachers. With computer-based testing, more constructs can be measured than by paper-pencil testing. This applies, for example, for measuring multimedia learning competencies, as videos and animations can be integrated into tests. The additional collection of intermediate data, such as response times and mouse movements, is also possible as well as adaptive testing according to students' capabilities (Csapó et al., 2014).

A disadvantage of computer-based testing is that investment in testing environments is expensive. The development of tests is time-consuming and needs careful planning. Some item formats that are used in paper-pencil testing cannot be implemented in computer-based testing due to technical reasons. Thus, the comparability between the two testing modes might in some cases be reduced. During online testing, problems with the internet connection might occur. Participants' computer literacy might also have an impact on their test performances (Kyllonen, 2009).

Comparison of computer-based and paper-pencil performance testing: empirical evidence

Below, we will give a broad overview of recent research on the comparison of computer-based testing with paper-pencil testing.

Jerrim (2016) analysed data of the PISA 2012 assessment in mathematics across 32 countries. The analyses yielded inconsistent results: In 11 countries (e.g. Hongkong), student performance was better in the paper-pencil assessment than in the computer-based assessment. In 13 countries (e.g. Brazil), the opposite pattern of results was found. Jerrim et al. (2018) investigated the mode effects in the PISA 2015 test scores in Germany, Sweden and Ireland. The authors found consistent evidence across countries and across the domains of mathematics and reading that students perform worse on the computer versions of the test. A further consideration of the mode effects for each item revealed that there were only small differences between computer-based testing and paper-pencil testing for 60 to 70% of the implemented test items.

Ackerman and Lauterman (2012) reported that the test performance of participants who had read expository texts either from screen or from paper did not differ significantly under a fixed study time. Under self-regulated study time, there was a mode effect disfavouring computer-based testing. Eyre (2017) investigated the equivalency of test scores of the computerized and the paper-pencil PAT reading comprehension test in New Zealand. Students' test scores were up to four scale points lower in the computerized version of the test. In the study of Dahan Golan et al. (2018), fifth- and sixth-graders achieved significantly higher scores in paper-based than in computer-based reading comprehension tests on narrative and expository texts. The results of the Norwegian study of Støle et al. (2020) also revealed that fifth-graders' average test performance in reading comprehension was lower in the computerized than in the paper-pencil version of the test.

In the study of Guimarães et al. (2017), students took either a computerized or a paper-pencil test on clinical anatomy. In the first testing session, students' test performance was significantly lower in the computer-based test than in the paper-pencil test. In the second testing session, the differences between the two testing modes were not significant. Prisacari (2017) investigated potential mode effects between a computerized and a paper-pencil test in

the domain of general chemistry in a sample of undergraduate students. There were no significant differences in the students' test performance between the two testing modes.

Potential impact of student characteristics on the mode effect

Gender

Recent research has yielded inconsistent evidence concerning gender differences in ICT literacy (for an overview, see Gnambs, 2021). In the German study of Gnambs (2021), 18-year-old boys had a slightly better ICT literacy than girls. In a meta-analysis, Siddiq and Scherer (2019) showed that girls' ICT literacy was significantly higher than that of boys. Van Deursen and van Diepen (2013) found no gender differences in information and strategic Internet skills. However, boys seem to have higher self-efficacy and more confidence in their higher-level ICT skills than girls (Cai et al., 2017; Fraillon et al., 2014), although there is evidence that this gender gap is also closing (Gnambs, 2021). The studies of Støle et al. (2020) and of Jerrim et al. (2018) comparing computer-based assessment with paper-pencil assessment showed no significant interactions between the testing mode and students' gender for the majority of participants. In the study of Jerrim (2016), boys outperformed girls in a mathematics assessment. This advantage was bigger for computer-based testing. Due to these inconsistent findings, it is hard to draw clear conclusions about the impact of students' gender on their performance in computer-based and paper-pencil assessments.

Academic achievement

VERA computer-based testing was recently introduced in Germany. Students first have to adapt to the new testing environment and its requirements. Interacting with the testing environment might induce an additional cognitive load and, at least in the first testing phase, might reduce students' working memory capacity for reflecting on and answering items (e.g. Sweller, 2020). Therefore, students with a lower academic achievement who have more difficulties in conceptually solving tasks might struggle more to adapt to the new testing requirements. However, this line of reasoning is only partly supported by previous empirical evidence: The study of Eno (2011) showed that high-achieving students profited more from computer-based than from paper-pencil testing. Jerrim et al. (2018) did not find a significant interaction between the testing mode and test achievement. In the study of Støle et al. (2020), top-performing girls were actually disadvantaged by computer-based testing.

Mainly spoken language in everyday life

Students whose mainly spoken language in everyday life is not German often come from immigrant families. Because these students have fewer opportunities to speak German, it might be harder for them to read and understand German texts properly compared with other students (Weis et al., 2019). Furthermore, on average, immigrant families have a lower socio-economic status than German families (Weis et al., 2019), and immigrant students might be less likely to have a computer at their disposal at home than other students. Therefore, it might be more difficult for them to cope with the additional effort of adapting to the new testing environment in the VERA computer-based assessment of the subject German at least in the first testing phases. The studies of Eyre (2017) and of Hardcastle et al. (2017) showed that

students from ethnic minorities or students whose primary language was not the test language were more disadvantaged by computer-based assessment than other students.

Study aims and hypotheses

VERA computer-based tests were recently introduced in Germany. There is almost no empirical evidence about the comparability of the scores of the VERA computerized test with the scores of the VERA paper-pencil test.

Therefore, the first aim of this study is to examine whether the scores from the VERA computerized test are different from the scores of the VERA paper-pencil test.

German students probably first have to familiarize themselves with the new VERA testing situation and its requirements. In the current study, students were tested in the subject German in the domains reading and orthography. In the studies of Eyre (2017), Dahan Golan et al. (2018) and Støle et al. (2020), students achieved lower reading comprehension scores in computer-based assessment than in paper-pencil-assessment. Therefore, we make the following assumption:

Hypothesis 1: In VERA tests, students perform worse in the computer-based test than in the paper-pencil test in the subject German.

The second study aim is to investigate whether students' gender, academic achievement and mainly spoken language in everyday life moderate the difference between VERA computer-based testing scores and paper-pencil testing scores.

The results of previous research on gender differences in ICT literacy and on the impact of students' gender on their performance in computer-based tests and paper-pencil tests are inconsistent (e.g. Gnambs, 2021; Jerrim, 2016; Siddiq & Scherer, 2019; Støle et al., 2020). Therefore, no specific hypotheses are made concerning the impact of students' gender on the mode effect in the current study.

Due to the mixed empirical evidence concerning the impact of student academic achievement on student performance in computer-based testing and paper-pencil testing (Eno, 2011; Jerrim et al., 2018; Støle et al., 2020), no specific hypotheses are made either on this issue.

Students whose mainly spoken language in everyday life is not German often have more problems in reading and understanding German texts and might have fewer computers at their disposal at home (Weis et al., 2019). Therefore, adapting to the new testing environment might be more difficult for them than for other students. In the studies of Eyre (2017) and of Hardcastle et al. (2017), students from ethnic minorities or students whose primary language was not the test language were more disadvantaged by computer-based testing than other students. Based on these considerations, we make the following assumption:

Hypothesis 2: Students whose mainly spoken language is not German are more disadvantaged by VERA computer-based testing than students whose mainly spoken language is German.

The third study aim is to explore how consistent the mode effect is across different tasks and task groups in VERA tests. Therefore, no specific hypotheses are made.

Method

Study design and procedure

The study was conducted with eighth-graders in secondary schools in four German federal states in February / March 2019. In three of the four investigated federal states, participation in the VERA tests in the subject German was obligatory for schools. The computer-based tests were implemented in a pilot study in two federal states. In general, schools could volunteer to participate in the computer-based tests; the study therefore had a field experimental design.

Before the paper-pencil test, teachers had to register their classes at an online platform and had to characterize their students in terms of several sociodemographic variables. In addition, they were instructed to study supplementary materials about how to conduct the VERA tests. In most of the investigated federal states, the booklets were printed by a central agency so that teachers were relieved of this extra effort. The test took 90 min. The time limit for the orthographic tasks was 37–38 min (depending on the booklet); the time limit for the reading tasks was 40 min. There was a break of 5 min between the two test sections. At the beginning of the test, teachers read aloud the standardized instructions to the students. Within the time limits, students were allowed to turn back pages and correct their answers. After the tests, teachers had to correct all responses, following the guidelines in a provided manual. Then, they had to enter the scores at an online platform; they received an immediate report from the platform with the preliminary results.

Before the computer-based test, teachers also had to register their classes at an online platform and specify several sociodemographic variables of their students. They were instructed to study supplementary material about the testing procedure including a manual of about 25 pages. They could also watch some tutorial videos. In addition, they could familiarize students with the tasks on the computer before testing. The testing times were the same as for the paper-pencil testing, except for the introduction, which took 5 min longer. At the beginning of the test, teachers read aloud the introduction. Students were also given some additional hints about the test on the screen. Within the time limits, students were allowed to return to tasks to correct their solutions. They were also allowed to use scrap paper during the test. After the test, teachers corrected the tasks with an open-ended response format on the computer. Omitted tasks and multiple-choice items were scored automatically by the computer system. After having corrected the tasks, teachers received an immediate report with the preliminary results. The hardware implemented in the tests—that is, the screen resolution of the monitors, operating systems and browsers—varied across schools.

Measures

Measurement of students' performance

The students' performance was measured using a variety of German orthographic and reading tasks. These tasks were designed to assess the competencies students are supposed to have in the subject German according to German educational standards (Kultusministerkonferenz, 2016). The suitability of these tasks was verified in a paper-pencil field trial of 1942 participants, which also served as a basis for selecting items from a larger pool. A Rasch model was used to model the competencies in orthography and reading. A set of items in the field trial was used to estimate the difficulties of the items in accordance with a quantifiable

model of the German educational standards, which is scaled to a mean of 500 and a standard deviation of 100. In order to keep the link with these standards, the item parameters estimated in the field trial were also used to model the competencies in the actual assessments.

In the German secondary school system, there are different types of schools for students with different levels of academic achievement. Therefore, the VERA tasks were split into two different booklets. Booklet 1 contained easier tasks and was recommended for schools with students with a lower level of academic achievement. This booklet consisted of five orthographic task groups with 62 single tasks and of three reading task groups with 41 single tasks. Booklet 2 contained more difficult tasks and was recommended for schools with students with a higher level of academic achievement. This booklet consisted of five orthographic task groups with 54 single tasks and of three reading task groups with 38 single tasks. Nevertheless, 23 orthographic tasks and 11 reading tasks were used in both booklets. In addition, the orthographic tasks were presented before the reading tasks in both booklets. There was less variation in the instructional format of the reading tasks compared with the orthographic tasks. The reading tasks typically required students to read a text and solve items that had a multiple-choice or an open-ended response format. The genre of the text to be read was expository for one task group of Booklet 1 and for two task groups of Booklet 2. It was narrative for two task groups of Booklet 1 and for one task group of Booklet 2. In addition to these task formats, orthographic tasks also required students to put commas in the right places of a text, for example, or to put hyphens in the right places in a word in order to separate its syllables correctly.

In both testing modes, the font type and font size used in the booklets was mainly Arial 12. The tasks were presented only in black and white.

Measurement of students' characteristics

- Gender

When registering their classes at the online platform, teachers could select from a drop-down menu to enter whether the student was male or female.

- Academic achievement

As an indicator for the moderating impact of academic achievement on the difference between computer-based test scores and paper-pencil test scores, we compared the two main effects of computer-based testing for Booklet 1 and Booklet 2. As Booklet 1 is normally given to students with lower academic achievement, and Booklet 2 is given to students with higher academic achievement, this comparison is supposed to give indications of the effect of this potentially moderating variable.

- Mainly spoken language in everyday life

Some bilingual students mainly speak or hear other languages than German in their everyday life. Teachers put a cross in the corresponding box at the online platform for these students when registering their classes.

In addition, teachers were instructed to provide information about students' learning disorders, such as dyslexia or hyperactivity, by checking the corresponding boxes at the online

platform. They were also instructed to enter whether students had repeated a school year. We included these variables in the analyses as control variables, as they presumably also have an impact on students' test performance.

Sample

In the four investigated federal states, 42,115 students participated in both domains of the VERA tests. The students were from 2182 different classes and from 630 different schools. Within this sample, 37,836 students (89.8%) participated in the paper-pencil test, and 4279 students (10.2%) participated in the computer-based test.

We excluded several students and classes from the statistical analyses. Students were excluded if they had special education needs or if they had participated in only either reading or orthography. Schools that voluntarily participated in the VERA tests were also excluded. This mostly concerned private schools or schools that care for children with special needs. In addition, we excluded classes that had technical problems conducting the computer-based tests. After this filtering procedure, 37,839 students remained in the sample.

The sizes of the computer-based and the paper-pencil test samples were very heterogeneous. In order to reduce the size of the paper-pencil test sample and to make it more comparable with the computer-based test sample regarding several context variables, we used "Propensity Score Matching" (Ho et al., 2011). Classes from the same federal state that solved the same booklet (1 or 2) and that were working towards attaining comparable school degrees were grouped together. Within each of these groups, the matching procedure was used in order to select a sample of classes within which several class characteristics were balanced as best as possible between the two testing conditions. The considered characteristics were class size, gender proportions, the proportion of students who did not speak German properly or whose mainly spoken language was not German, the proportion of students with special educational needs and the proportion of students who had to repeat the year. After this matching procedure, the standardized mean differences of all context variables between the two conditions were smaller than .01.

Table 1 shows the sizes of the whole sample and of the sub-samples that were used for the statistical analyses.

Table 2 shows the frequencies of the investigated student characteristics in the computer-based test condition and in the paper-pencil test condition for both booklet sub-samples.

Table 2 shows that, in the sub-sample that completed Booklet 2, the number of students that had German as the non-dominant language, who were repeating Year 8 or had learning disorders, was rather low.

Table 1 Sample sizes of the whole sample and of the sub-samples that completed Booklet 1 and Booklet 2

	Computer-based testing			Paper-pencil testing			Total		
	Schools	Classes	Students	Schools	Classes	Students	Schools	Classes	Students
Whole sample	52	148	2318	113	148	2822	165	296	5140
Sub-sample Booklet 1	34	95	1420	70	95	1726	104	190	3146
Sub-sample Booklet 2	18	53	898	43	53	1096	61	106	1994

Table 2 Frequencies of the investigated student characteristics in the computer-based test condition and in the paper-pencil test condition for both booklet sub-samples

	Computer-based testing				Paper-pencil testing			
	Girls	GND	YR	LD	Girls	GND	YR	LD
Sub-sample Booklet 1	660 (46.5%)	99 (7.0%)	16 (1.1%)	47 (3.3%)	793 (45.9%)	122 (7.1%)	27 (1.6%)	55 (3.2%)
Sub-sample Booklet 2	451 (50.2%)	10 (1.1%)	5 (0.6%)	13 (1.4%)	574 (52.4%)	11 (1.0%)	7 (0.6%)	19 (1.7%)

Girls = frequency of girls as an indicator for sex distribution, *YR* year repetition, *GND* German as non-dominant language, *LD* learning disorder

Statistical analyses

In order to test the hypotheses concerning the main effect of computer-based testing and its interactions with student characteristics, we used multilevel modelling. On the first level, we examined the predictors "Mode", "Gender", "German as non-dominant language" and the control variables "Learning disorder" and "Year repetition". On the second level, we took into account the fact that the students belonged to different classes, but included no further predictors. The dependent variable was student performance as measured by the Rasch model. We calculated one model for each combination of booklet and domain. To estimate the coefficients, we used the REML-method in the packages "lme4" and "lmerTest" of the statistical software "R".

In order to investigate the mode effects on an item-by-item basis, we calculated the fraction of correct solutions for each item under both conditions.

Results

Tables 3, 4, 5, and 6 show the results of the multilevel analyses.

Table 3 shows that in the orthographic domain of Booklet 1 student performance in the VERA computer-based test was significantly lower than student performance in the VERA paper-pencil test. The effect size was about two-thirds of a standard deviation. Girls had significantly higher scores than boys, but this advantage was lower in the computer-based test.

Table 3 Results of the mode effect and its interactions with student characteristics (Orthography, Booklet 1)

	Coefficient	SE	*t*-value	*p*
Intercept	449.64	5.82	77.22	< .001***
CBT	− 64.33	8.39	− 7.66	< .001*** (one-tailed)
Sex	− 40.52	4.32	− 9.37	< .001***
GND	− 46.34	10.22	− 4.53	< .001***
YR	5.46	18.44	.30	.77
LD	− 61.41	12.46	− 4.93	< .001***
CBT × sex	13.27	6.44	2.06	.04*
CBT × GND	− 18.98	14.04	− 1.35	.09 (one-tailed)
CBT × YR	− 33.76	29.41	− 1.15	.26
CBT × LD	12.46	18.41	.68	.50

CBT computer-based testing, sex: girls = 0, boys = 1, *GND* German as non-dominant language, *YR* year repetition, *LD* learning disorder

Table 4 Results of the mode effect and its interactions with student characteristics (Reading, Booklet 1)

	Coefficient	SE	*t*-value	*p*
Intercept	467.82	6.25	74.89	<.001***
CBT	− 60.79	9.00	− 6.75	< .001*** (one-tailed)
Sex	− 23.04	4.62	− 4.99	< .001***
GND	− 77.47	10.92	− 7.09	<.001***
YR	− 4.74	19.69	− .24	.81
LD	− 32.15	13.30	− 2.42	.02*
CBT × sex	14.12	6.88	2.05	.04*
CBT × GND	7.72	15.00	.52	.61
CBT × YR	2.79	31.41	.09	.93
CBT × LD	32.95	19.66	1.68	.09

CBT computer-based testing, sex: girls = 0, boys = 1, *GND* German as non-dominant language, *YR* year repetition, *LD* learning disorder

Students who did not mainly speak German in their everyday life performed non-significantly worse in the VERA computer-based test than other students.

In the reading domain of Booklet 1, student performance in the VERA computer-based test was again significantly lower than in the VERA paper-pencil test (see Table 4). This difference was again about 60 points. Similarly to in the orthographic domain, girls performed significantly better than boys, but this difference was smaller in the computer-based test. There was no interaction between testing mode and students' mainly spoken language in everyday life.

Table 5 shows that student performance in the orthographic domain of Booklet 2 was more than one standard deviation (120 points) higher than student performance in Booklet 1. Again, students performed significantly worse in the VERA computer-based test than in the paper-pencil test. The difference in score was about 33 points and therefore only half as high as for Booklet 1. Girls performed significantly better than boys. Conversely to Booklet 1, there was no interaction between testing mode and students' gender. Students with German as the non-dominant language were significantly disadvantaged in the VERA computer-based test compared with other students.

In the reading domain of Booklet 2, the students' performance was again more than one standard deviation higher than in Booklet 1 (see Table 6). Students' performance in the VERA computer-based test was again significantly lower than in the paper-pencil test, and the effect size was about the same as in the orthographic domain (28 points). Girls scored significantly higher than boys and seemed to perform slightly better in the VERA computer-based test.

Table 5 Results of the mode effect and its interactions with student characteristics (Orthography, Booklet 2)

	Coefficient	SE	*t*-value	*p*
Intercept	569.80	6.67	85.39	< .001***
CBT	− 32.88	9.68	− 3.40	< .001*** (one-tailed)
Sex	− 20.40	4.37	− 4.67	< .001***
GND	− 41.71	21.77	− 1.92	.06
YR	− 11.56	26.92	− .43	.67
LD	− 82.64	16.40	− 5.04	< .001***
CBT × sex	− 8.99	6.52	− 1.38	.17
CBT × GND	− 71.27	31.63	− 2.25	.01* (one-tailed)
CBT × YR	− 89.21	41.75	− 2.14	.03*
CBT × LD	43.93	25.47	1.73	.09

CBT computer-based testing, sex: girls = 0, boys = 1, *GND* German as non-dominant language, *YR* year repetition, *LD* learning disorder

Table 6 Results of the mode effect and its interactions with student characteristics (Reading, Booklet 2)

	Coefficient	SE	*t*-value	*p*
Intercept	579.41	7.60	76.28	< .001***
CBT	− 28.33	11.07	− 2.56	.005** (one-tailed)
Sex	− 14.12	5.70	− 2.48	.01*
GND	− 88.30	28.37	− 3.11	.002**
YR	− 66.42	35.10	− 1.89	.06
LD	− 14.43	21.38	− .68	.50
CBT × sex	− 14.66	8.49	− 1.73	.09
CBT × GND	− 72.67	41.21	− 1.76	.04* (one-tailed)
CBT × YR	− 29.11	54.41	− .54	.59
CBT × LD	− 32.09	33.23	− .97	.33

CBT computer-based testing, sex: girls = 0, boys = 1, *GND* German as non-dominant language, *YR* year repetition, *LD* learning disorder

Students with German as the non-dominant language were strongly disadvantaged in the VERA computer-based test.

All of the above results hinge on the assumption that, apart from the testing mode, the computer-based tasks and the paper-pencil tasks measure the same competencies in an analogous way. This was addressed by our third study aim, namely to determine whether the size of the mode effect varied across tasks or task groups. If it did, the psychometric model used to model the competencies could no longer be assumed to hold.

The linear correlation between the item solution rates under the paper-pencil and computer-based condition seemed to be sufficiently high for both reading blocks (r = .97) and for orthography in Booklet 1 (r = .95). For orthography in Booklet 2, however, the correlation was only r = .91, indicating important deviations in item difficulties. The mode effect on the task solution rates is explored further in Figs. 1, 2, 3, and 4. The long vertical lines in the figures represent different task groups.

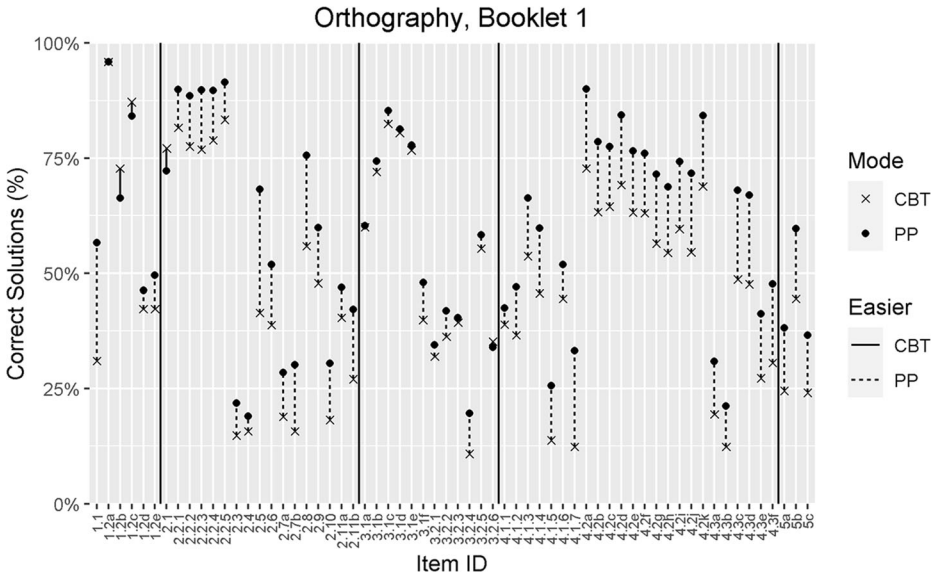

Fig. 1 Mode effects for each item in the orthographic domain of Booklet 1

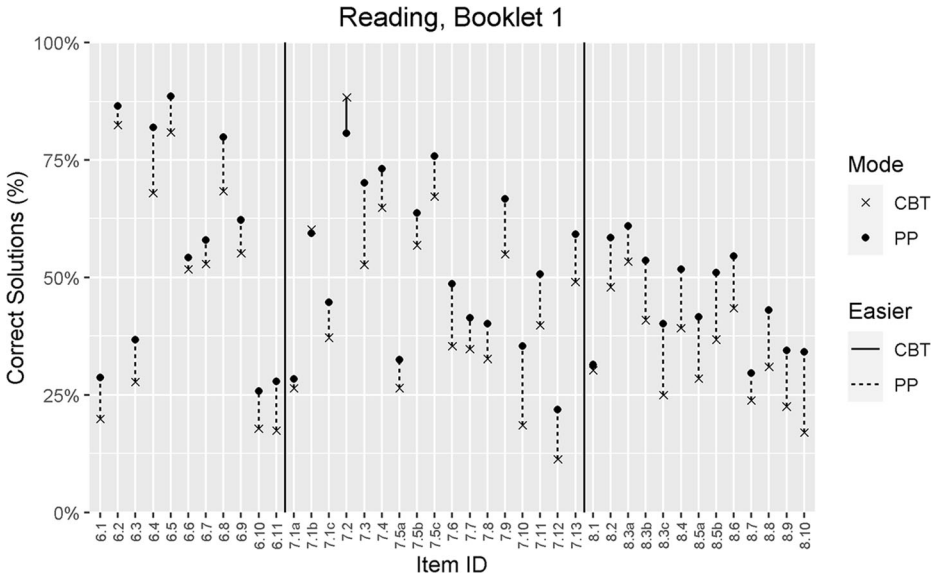

Fig. 2 Mode effects for each item in the reading domain of Booklet 1

Figures 1, 2, 3, and 4 show that the correct solution rates were lower for VERA computer-based testing compared with VERA paper-pencil testing for 175 out of 195 items (89.7%). The size of the mode effect seemed to vary across task groups. This variation was stronger in the orthographic than in the reading domain. A close inspection and explorative analysis did not reveal any task characteristics that consistently predicted the size of the mode effect disfavouring computer-based testing. On average, small mode effects were found for Task Group 3 of the orthographic domain in both booklets (Figs. 1 and 3). These tasks had a clear

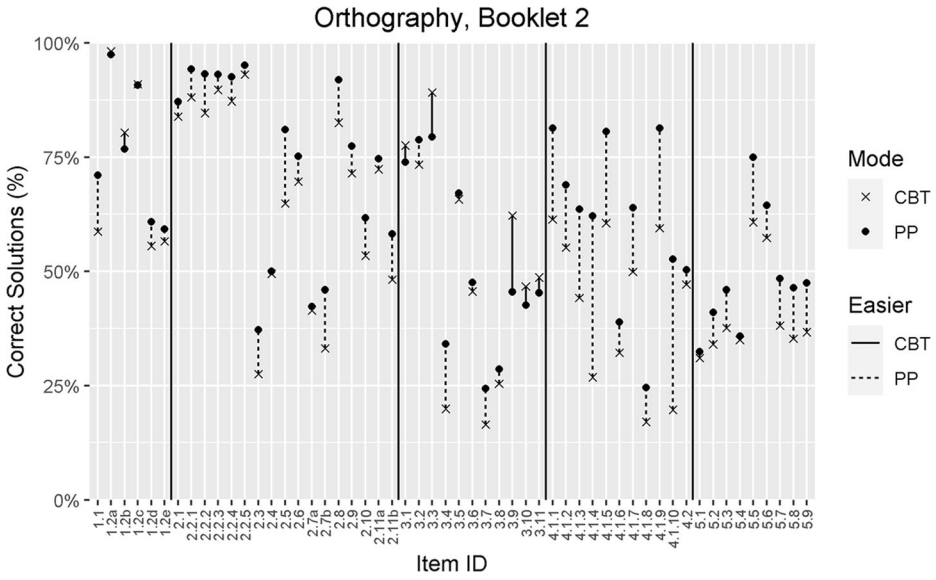

Fig. 3 Mode effects for each item in the orthographic domain of Booklet 2

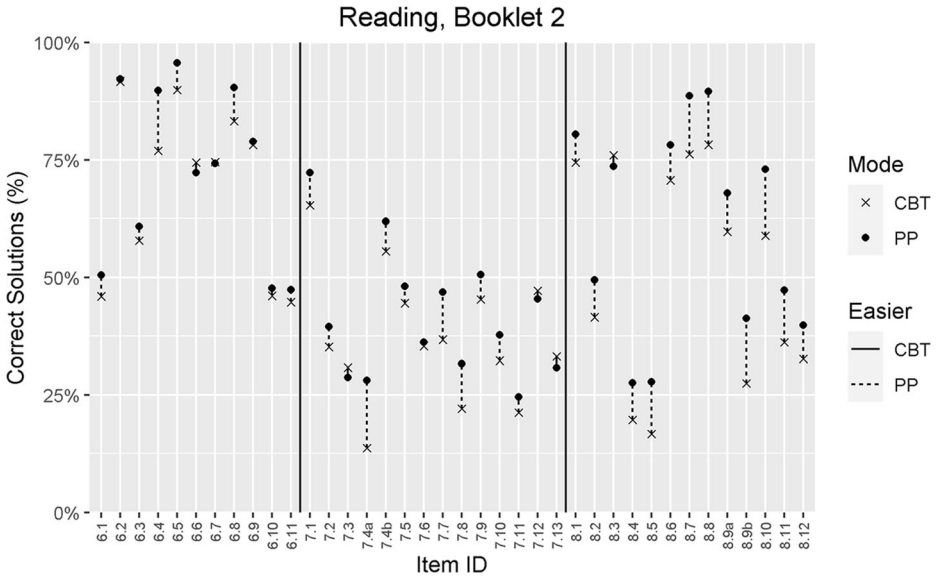

Fig. 4 Mode effects for each item in the reading domain of Booklet 2

instructional format requiring students to put commas in the right places of a text which involved using the mouse in the computer-based test. On average, large mode effects were found for Task Group 4 of the orthographic domain in Booklet 2 (Fig. 3). These tasks had a complex instructional format, requiring students to take several actions; they had to both check boxes using the mouse and type answers using the keyboard in order to solve the tasks in the computer-based test. Students might have also had to scroll back up the screen to re-read the instructions. There were also variations of the mode effect within task groups which could not be explained as the tasks had the same instructional format.

Discussion

Summary of results and conclusions

The results of the current study revealed that student performance in the German VERA tests in the year 2019 was significantly lower in the computer-based test than in the paper-pencil test in the subject German. This result applied to the testing domains orthography and reading and for both test booklets, which were completed by different student samples with varying levels of academic achievement. The effect sizes were rather high: For Booklet 1, the difference between the computer-based and paper-pencil test scores was about .60 standard deviations; for Booklet 2 it was about .30 standard deviations. Therefore, Hypothesis 1 was confirmed: Students performed worse in the VERA computer-based test than in the paper-pencil test. The results are consistent with the results of the studies of Eyre (2017), Dahan Golan et al. (2018) and Støle et al. (2020), in which students achieved lower reading comprehension scores in computer-based tests than in paper-pencil tests. The poorer student performance in the computer-based test might have been because students were not yet accustomed to this kind of testing. Most of the tests and exams in the classroom are still written in the paper-pencil format. Computer-based testing offers several

advantages but also involves challenges, such as typing long phrases using a keyboard or scrolling to find a specific piece of information to solve a task. Students first have to become used to meeting these demands before they can fully concentrate on understanding and solving the tasks conceptually. In addition, teachers also have to become accustomed to conducting computer-based performance tests. Teachers supervising the computer-based testing condition were possibly more insecure and had more technical difficulties than the teachers in the paper-pencil testing condition, which could have negatively affected the atmosphere during the test.

The results regarding the interaction between the testing mode and students' gender were inconsistent. In this study, the girls outperformed the boys. This advantage decreased significantly in the computer-based condition in both domains of Booklet 1, but it non-significantly increased in both domains of Booklet 2. These ambiguous results are in line with the inconsistent findings of recent research on gender differences in ICT literacy, ICT self-efficacy and the impact of students' gender on the mode effect between computer-based testing and paper-pencil testing (e.g. Gnambs, 2021; Jerrim, 2016; Støle et al., 2020).

As a descriptive indicator for the moderating effect of students' academic achievement, we compared the two mode effects between computer-based and paper-pencil testing for the different VERA test booklets. Teachers were instructed to use Booklet 1 to test classes with a lower level of academic achievement. Booklet 2 was recommended for higher-achieving classes. The students' test scores were more than one standard deviation higher in both domains in Booklet 2 compared with Booklet 1. The analyses showed that the disadvantage of the computer-based testing condition for Booklet 2 was only half as high as for Booklet 1. Thus, students with lower academic achievement seemed to be more disadvantaged by computer-based testing than students with higher academic achievement in the VERA performance tests. This result is in line with the findings of Eno (2011). It might also support the argumentation that low-achieving students have more problems with conceptually solving tasks and might therefore be more easily overburdened by the additional effort of adapting to the new testing environment than high-achieving students.

Students who did not mainly speak German in their everyday life generally performed worse than other students. There was some evidence that they were disadvantaged by VERA computer-based testing but the effects were inconclusive for Booklet 1. Furthermore, only few children with German as the non-dominant language were tested with the second booklet, where the effects were much larger. Thus, the second hypothesis, namely that students whose mainly spoken language is not German are disadvantaged by VERA computer-based testing, could only partly be confirmed. Nevertheless, students who had fewer opportunities to speak German in their everyday lives tended to have more difficulties with VERA computer-based testing in the subject German than other students. Students who are less accustomed to speaking German might have more problems solving the German orthographic and reading tasks and, therefore, be overburdened more easily by the new testing requirements. In addition, these students often come from immigrant families that have a lower socio-economic status on average (Weis et al., 2019) and might therefore be less likely to have a computer at their disposal at home.

Even though it seems that some students might have been disadvantaged by VERA computer-based testing in 2019, it should be stated that, once familiarized with this new testing medium, students might profit from computer-based testing in the long term. For example, in computer-based testing, students can listen to pre-recorded instructions which are read aloud (Csapó et al., 2014), thus reducing irrelevant reading comprehension difficulties, e.g. in mathematics assessments. In addition, the implementation of new item formats, such as videos, and adaptive testing is possible in computer-based tests (Csapó et al., 2014).

It has to be noted that all of the above results have to be regarded with some caution because the mode effect varied across tasks. This indicates that the change of testing mode did not result in an identical shift in item difficulty, but affected at least some of the items in other ways. Thus, it cannot be assumed that the computer-based version of the assessments measures the target competencies in strictly the same way as the paper-pencil test. As discussed below, this problem was more pronounced for orthography than it was for reading.

The third study aim was to investigate the variance of the mode effect across tasks. The explorative results of the current study revealed that the correct solution rates were lower for the computer-based test compared with the paper-pencil test for 175 out of 195 items. The size of the mode effect varied both across tasks and task groups. The variation across task groups was stronger in the orthographic than in the reading domain. This might have been due to the different instructional formats of the orthographic task groups, whereas the instructional formats of the reading task groups were quite similar to each other. It remains unclear which task characteristics lead to stronger mode effects, as a preliminary explorative analysis did not reveal any consistent results. This is in line with the results of Poggio et al. (2005) who did not find any clear factors accounting for items that are more difficult in computer-based testing, despite having conducted a close inspection of these items. In the current study, small mode effects disfavouring computer-based testing were found for Task Group 3 in the orthographic domain of Booklets 1 and 2. These tasks had a clear instructional format, requiring students to put commas in the right places of a text using the mouse in computer-based testing. They did not require scrolling. The largest mode effects were found for Task Group 4, also in the orthographic domain of Booklet 2. These tasks had a complex instructional format, requiring students to give answers both by checking boxes with a mouse and by typing out answers on a keyboard in the computer-based test. In order to re-read the instructions, students had to scroll up the screen. One first assumption might be that simple tasks without scrolling and where only the use of the mouse is required might lead to smaller mode effects disfavouring computer-based testing. Complex tasks that require scrolling and the use of both mouse and keyboard might cause larger mode effects. However, further studies analysing the impact of task characteristics on mode effects more deeply are needed to confirm these assumptions.

Implications

In sum, the results of the current study revealed that there was a stronger mode effect disfavouring the computer-based version of VERA tests compared with the paper-pencil version. This mode effect was consistent across different testing domains and test booklets. This implies that the computer-based test scores in the VERA tests in the year 2019 in the subject German were not directly comparable with the paper-pencil test scores from previous years or from other federal states. The equivalency of computer-based test scores and paper-pencil test scores was not yet given.

Students' disadvantage in computer-based performance testing might decrease in the coming years as they become more and more accustomed to this new testing medium and its requirements. Teachers might also acquire a routine and become more confident in conducting computer-based performance tests. Nevertheless, this development could be accelerated by integrating computer-based tests and examinations into normal classroom routines more frequently. In addition, students could be explicitly instructed on how to meet the requirements of computer-based testing. Teachers should become more professionalized in developing and conducting computer-based test formats. Head teachers might motivate them to use this new testing format.

As a further supportive measure, head teachers could encourage collaboration among teachers regarding the implementation of computer-based testing. Finally, a teacher with technical expertise could be assigned to help the other teachers with conducting computer-based testing.

Students with lower academic achievement and with German as the non-dominant language tended to achieve lower scores in VERA computer-based testing. Teachers might prepare these students more intensely before and support them more during the performance tests, at least during the first testing cycles.

In principle, teachers have the opportunity to familiarize students with tasks of the VERA computer-based tests. They could be more encouraged by head teachers or by instructions in the supplementary material to make use of this offer.

Methodological constraints

A major constraint was the fact that the assessments were designed and psychometrically modelled solely with paper-pencil testing in mind. Thus, the tasks had to be adopted to a computer-based format.

The number of students whose dominant language was not German was rather low for Booklet 2.

Another deficit of the study was its field experimental design. Schools volunteered to participate in computer-based testing. Thus, better equipped and more innovative schools might have conducted computer-based performance testing.

Finally, it remains unclear as to whether and to what extent the higher paper-pencil test scores compared with computer-based test scores were caused by teachers correcting the tasks less strictly than the computer system.

Research perspectives

Further studies on the mode effect between computer-based and paper-pencil performance testing should be conducted with experimental designs and large sample sizes. Drawing randomized samples from the whole school population might ensure that the samples, especially in computer-based tests, are more representative.

In the current study, we analysed the impact of task characteristics on the mode effect only exploratively and cannot draw any clear conclusions yet. Therefore, further studies involving quantitative regression modelling and qualitative task analyses should investigate the impact of task characteristics on the mode effect more deeply.

Finally, previous research has shown that the mode effects between computer-based testing and paper-pencil testing are often inconsistent and seem to depend on different factors. The results of the current study are valid for VERA performance tests in the subject German in four German federal states in the year 2019. They apply to the specific testing conditions under which the tests were conducted. The generalizability of the results to other federal states and countries, subjects, years and testing conditions remains unclear. Therefore, the mode effect between computer-based and paper-pencil performance testing needs to be investigated further. Research into different subjects and in other federal states and countries under different testing conditions might lead to a better understanding of the mode effect and its moderating factors.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Ministries of Education of four German federal states. There was no impact of the Ministries on data interpretation and the writing of this article.

Declarations

Conflict of interest The authors declare no competing interests.

References

Ackerman, R., & Lauterman, T. (2012). Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. *Computers in Human Behavior, 28*(5), 1816–1828. https://doi.org/10.1016/j.chb.2012.04.023.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6*(9) Retrieved March 18, 2020 from https://ejournals.bc.edu/index.php/jtla/article/view/1639.

Cai, Z., Fan, X., & Du, J. (2017). Gender and attitudes toward technology use: A meta-analysis. *Computers & Education, 105*, 1–13. https://doi.org/10.1016/j.compedu.2016.11.003.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology, 33*(5), 593–602. https://doi.org/10.1111/1467-8535.00294.

Csapó, B., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school readiness and early reasoning. *Journal of Educational Psychology, 106*(3), 639–650. https://doi.org/10.1037/a0035756.

Dahan Golan, D., Barzillai, M., & Katzir, T. (2018). The effect of presentation mode on children's reading preferences, performance, and self-evaluations. *Computers & Education, 126*, 346–358. https://doi.org/10.1016/j.compedu.2018.08.001.

Ejim, S. (2017). *Overview of computer-based tests*. Research proposal. https://doi.org/10.13140/RG.2.2.32040.88326.

Eno, L. (2011). *Comparing the reading performance of high-achieving adolescents: Computer-based testing versus paper/pencil*. Dissertation. Seton Hall University. Retrieved August 10, 2020 from https://search.proquest.com/docview/1018431561.

Eyre, J. (2017). On or off screen reading in a digital world. *Assessment News Set, 1*, 53–58. https://doi.org/10.18296/set.0072.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). Students' use of and engagement with ICT at home and school. In J. Fraillon, J. Ainley, W. Schulz, T. Friedman, & E. Gebhardt (Eds.), *Preparing for life in a digital age* (pp. 125–166). Springer. https://doi.org/10.1007/978-3-319-14222-7.

Gnambs, T. (2021). The development of gender differences in information and communication technology (ICT) literacy in middle adolescence. *Computers in Human Behavior, 114*, 106533. https://doi.org/10.1016/j.chb.2020.106533.

Guimarães, B., Ribeiro, J., Cruz, B., Ferreira, A., Alves, H., Cruz Correia, R., Madeira, M. D., & Ferreira, M. A. (2017). Performance equivalency between computer-based and traditional pen-and-paper assessment: A case study in clinical anatomy. *Anatomical Sciences Education, 11*(2), 124-136. https://doi.org/10.1002/ase.1720

Hardcastle, J., Herrmann-Abell, C. F., & DeBoer, G. E. (2017). *Comparing student performance on paper-and-pencil and computer-based-tests.* Paper presented at the 2017 AERA Annual Meeting, San Antonio, TX. Retrieved December 8, 2020 from https://files.eric.ed.gov/fulltext/ED574099.pdf

Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts [Quality of classroom practices and teachers' professionalism. Diagnosis, evaluation and improvement of lessons]* (4th ed.). Seelze: Klett-Kallmeyer.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*(8) Retrieved August 12, 2020 from http://www.jstatsoft.org/.

Jerrim, J. (2016). PISA 2012: how do results for the paper and computer tests compare? *Assessment in Education: Principles. Policy & Practice, 23*(4), 495–518. https://doi.org/10.1080/0969594X.2016.1147420.

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: how big is the 'mode effect' and what has been done about it? *Oxford Review of Education, 44*(4), 476–493. https://doi.org/10.1080/03054985.2018.1430025.

Kultusministerkonferenz. (2016). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitorin [Strategic approach of the Standing Conference of the Ministers of Education and Cultural Affairs on educational monitoring]*. Wolters Kluwer & KMK.

Kyllonen, P. C. (2009). New constructs, methods, & directions for computer-based assessment. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 151–156). Office for official publications of the European communities. https://doi.org/10.2788/60083.

Poggio, J., Glasnapp, D. R., Yang, X., & 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. *Journal of Technology, Learning, and Assessment, 3*(6). Retrieved March 18, 2020 from https://ejournals.bc.edu/index.php/jtla/article/view/1659.

Prisacari, A. A. (2017). *Measuring the testing mode in general chemistry: The effect of computer versus paper mode on test performance, cognitive load, and scratch paper*. Dissertation. Iowa State University. Retrieved December 14, 2020 from https://www.proquest.com/docview/1918635382.

Siddiq, F., & Scherer, R. (2019). Is there a gender gap? A meta-analysis of the gender differences in students' ICT literacy. *Educational Research Review, 27*, 205–217. https://doi.org/10.1016/j.edurev.2019.03.007.

Stecher, B. M. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 79–100). RAND Corporation.

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

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development, 68*(1), 1–16. https://doi.org/10.1007/s11423-019-09701-3.

Thompson, N. A., & Weiss, D. J. (2009). Computerized and adaptive testing in educational assessment. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 127–133). Office for official publications of the European communities. https://doi.org/10.2788/60083.

Van Deursen, A. J. A. M., & van Diepen, S. (2013). Information and strategic Internet skills of secondary students: A performance test. *Computers & Education, 63*, 218–226. https://doi.org/10.1016/j.compedu.2012.12.007.

Weis, M., Müller, K., Mang, J., Heine, J.-H., Mahler, N., & Reiss, K. (2019). Soziale Herkunft, Zuwanderungshintergrund und Lesekompetenz [Social and migrational backgound and reading competence]. In K. Reiss, M. Weis, E. Klieme, & O. Köller (Eds.), *PISA 2018. Grundbildung im internationalen Vergleich [PISA 2018. Education considered in an internationally comparative context.]* (pp. 129–162). Waxmann. https://doi.org/10.31244/9783830991007.

Dr. Inga Wagner. Centre for Educational Research, University of Koblenz-Landau Buergerstrasse 23 D-76829 Landau Germany. E-mail: iwagner@zepf.uni-landau.de

Current themes of research:

School inspection. Performance tests as an external evaluation instrument in school systems. Text and picture comprehension.

Most relevant publications in the field of Psychology of Education:

Wagner, I. (2020). Effectiveness and perceived usefulness of follow-up classroom observations after school inspections in Northern Germany. Studies in Educational Evaluation, 67, 100913. https://doi.org/10.1016/j.stueduc.2020.100913

Zhao, F., Schnotz, W., Wagner, I., & Gaschler, R. (2020). Texts and Pictures Serve Different Functions in Conjoint Mental Model Construction and Adaptation. *Memory & Cognition, 48*, 69-82. https://doi.org/10.3758/s13421-019-00962-0

Wagner, I., Hosenfeld, I., & Zimmer-Müller, M. (2019). Vergleichende Analyse der Zusammenhänge von Akzeptanz, Auseinandersetzung mit und Nutzung von Ergebnissen bei Vergleichsarbeiten und Schulinspektionen. [Comparative analysis of the relationships between acceptance, reflection and use of results of school inspections and performance tests]. *Psychologie in Erziehung und Unterricht.* https://doi.org/10.2378/peu2019.art22d

Schnotz, W., & Wagner, I. (2018). Construction and elaboration of mental models by students of different grades and school tiers through strategic conjoint processing of text and pictures. *Journal of Educational Psychology, 110*(6), 850-863. https://doi.org/10.1037/edu0000246

Schnotz, W., Wagner, I., Ullrich, M., Horz, H., & McElvany, N. (2017). Development of students' text-picture integration and reading competence across grades 5 to 7 in a three-tier secondary school system: a longitudinal study. *Contemporary Educational Psychology, 51*, 152-169. https://doi.org/10.1016/j.cedpsych.2017.06.003

Wagner, I., & Schnotz, W. (2017). Learning from static and dynamic visualizations: What kind of questions should be asked? In R. K. Lowe & R. Ploetzner (Eds.), *Learning from dynamic visualizations: Innovations in research and application* (pp. 69-91). New York: Springer. https://doi.org/10.1007/978-3-319-56204-9

Philipp Loesche. Centre for Educational Research, University of Koblenz-Landau Buergerstrasse 23 D-76829 Landau Germany. E-mail: loesche@zepf.uni-landau.de

Current themes of research:

Gender differences in scholastic achievement. Neural networks

Most relevant publications in the field of Psychology of Education:

Loesche, P. M. (2019). Estimating the true extent of gender differences in scholastic achievement: A neural network approach. *Intelligence, 77*. https://doi.org/10.1016/j.intell.2019.101398.

Steven Bißantz. Centre for Educational Research, University of Koblenz-Landau Buergerstrasse 23 D-76829 Landau Germany. E-mail: bissantz@zepf.uni-landau.de

Current themes of research:

Statistical modelling.

Most relevant publications in the field of Psychology of Education:

No previous publications.

Affiliations

Inga Wagner[1] · Philipp Loesche[1] · Steven Bißantz[1]

Philipp Loesche
loesche@zepf.uni-landau.de

Steven Bißantz
bissantz@zepf.uni-landau.de

[1] Centre for Educational Research, University of Koblenz-Landau, Buergerstrasse 23, D-76829 Landau, Germany