British Journal of Educational Technology *Vol 0 No 0 2020* 1–19
doi:10.1111/bjet.12914

Tablet assessment in primary education: Are there performance differences between TIMSS' paper-and-pencil test and tablet test among Dutch grade-four students?

Eva Hamhuis, Cees Glas and Martina Meelissen

Eva Hamhuis is a researcher in the Department of Research Methodology, Measurement and Data Analysis at the University of Twente. She was member of the Dutch research team of TIMSS 2019. Her research interests include large-scale assessments and the role of ICT in education. Cees Glas is a professor in the Department of Research Methodology, Measurement and Data Analysis at the University of Twente. His research interests include educational measurement, item response theory and Bayesian statistical modelling. Martina Meelissen is a senior researcher in the Department of Research Methodology, Measurement and Data Analysis at the University of Twente. She is the national research coordinator of TIMSS and PISA. Her, research interests include large-scale assessments, gender differences and STEM in education. Address for correspondence: Eva Hamhuis, Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: e.r.hamhuis@utwente.nl

Abstract
Over the last two decades, the educational use of digital devices, including digital assessments, has become a regular feature of teaching in primary education in the Netherlands. However, researchers have not reached a consensus about the so-called "mode effect," which refers to the possible impact of using computer-based tests (CBT) instead of paper-and-pencil-based tests (PBT) to assess student performance. Some researchers suggest that the occurrence of a mode effect might be related to the type of device used, the subject being assessed and the characteristics of both the test and the students taking the test. The international TIMSS 2019 Equivalence Study offered the opportunity to explore possible performance differences between a PBT and a tablet assessment in mathematics and science among Dutch primary school students. In the spring of 2017, the TIMSS PBT and tablet test were administered to 532 grade-four Dutch students. Item response theory was used to explore potential mode effects. This exploration revealed no significant differences in the student ability scales between the paper and the tablet tests for mathematics and science. Also, no systematic mode effects were found for the items with high reading demand. A marginal difference was found for girls outperforming boys on the TIMSS tablet test, although no gender differences in achievement were found for the PBT.

Introduction

The so-called "app generation" is growing up with the latest information and communication technology (ICT), both in and outside school. Of all the different devices available (computers, laptops, tablets, chrome books, mobile phones), the use of tablets has recorded the fastest growth in educational settings in the United States (Poll, 2015). Since the introduction of the first Apple iPad tablet in 2010, followed by Google Android-based tablets, the use of tablets in Dutch primary schools has also grown rapidly (Brummelhuis & Binda, 2017). In Dutch primary schools, tablet

Practitioner Notes

Previous knowledge about this topic

- Many studies have investigated the test mode differences between CBT and PBT, but their results are divergent and inconclusive.
- Factor familiarity with the test device has an important impact on test performance.
- The difference between reading on paper or on a tablet is an important factor. Many studies demonstrate that when reading on a screen, text comprehension is perceived to be more difficult.

What this paper contributes

- This equivalence study shows that CBT is highly comparable to PBT, indicating that both test modes are suitable for testing the conceptual knowledge of mathematics and science and are comparable for assessing TIMSS in the Netherlands.
- When assessed on a tablet, Dutch grade-four girls slightly outperform boys in mathematics.

Implications for practice and/or policy

- Within the Netherlands, the tablet test could be a good alternative to the traditional PBT.
- Digital testing enables investigations on the differences between boys and girls in terms of response behaviour, such as response time by analysing the log files of the assessment.

computers are not only frequently used for instruction and learning, they are also used for student assessments (Faber & Visscher, 2016).

The educational use of computers in schools has resulted in a great deal of interest from educational researchers regarding the benefits and limitations of technology use in the classroom (Shute & Rahimi, 2017). There has been extensive research on the differences in student achievement from traditional paper-and-pencil-based tests (PBT) versus computer-based tests (CBT) (Noyes & Garland, 2008). However, the results of this research are inconclusive, and the occurrence of the so-called "mode effect" seems to be related to specific factors, such as the subject matter tested, students' characteristics, the user-friendliness of the testing device and the specific characteristics of the assessment (eg, Dadey, Lyons, & DePascale, 2018; Jeong, 2014; Jerrim, 2016). Furthermore, most of the research has focused on the use of computers and less on the use of tablets for assessments.

Using the Dutch data from the Trends in International Mathematics and Science Study (TIMSS) 2019 Equivalence Study, this study explores differences in primary school student achievement in mathematics and science by comparing PBT and a tablet test in relation to gender and the reading demand of mathematics items. TIMSS is an international large-scale assessment study that began in 1995 and is conducted every 4 years. The Equivalence Study is part of the TIMSS 2019 cycle, in which the administration of the TIMSS assessment transformed from a PBT to a CBT in about half of the participating countries.

Review of the literature

A substantial number of studies from the previous two decades focused on the effects of using digital testing to assess student performance. This "mode effect" refers to the likelihood of differential

student performance due to differences in how items are presented in PBT versus CBT (Wang, Jiao, Young, Brooks, & Olson, 2007). The results of these studies were mixed, as differences in student achievement were often found to favour PBT, with some studies reporting higher achievement in CBT or no performance differences at all (Bennett *et al.*, 2008; Clariana & Wallace, 2002; Jerrim, 2016; Nardi & Ranieri, 2018; Noyes & Garland, 2008; Russell, Goldberg, & O'Connor, 2003; Wang *et al.*, 2007). Research suggests that the occurrence of mode effects depends on the characteristics of the students taking the test (eg, gender, socio-economic status or experience with the device) as well as the test characteristics, such as the subject matter being tested, the user-friendliness of the testing device, and how the test items are presented.

Student characteristics

Previous research has shown that female students perform slightly better on PBT than on CBT (eg, Gallagher, Bridgeman, & Cahalan, 2002). According to Jeong (2014), the discrepancies between CBT and PBT scores are a function of gender and the subject matter being tested. Jeong also found that the CBT scores of girls in primary education in Korea were significantly lower than their PBT scores, especially in mathematics and science tests. Girls and boys had comparable PBT scores in science and mathematics, but girls' CBT scores were lower than those of boys. An exploration of the data of the Programme for International Student Assessment (PISA) 2012 showed that the overall gender gap in mathematics (boys outperforming girls) is slightly wider for CBT compared to PBT (Jerrim, 2016). In the data, 32 countries or educational systems administrated both a PBT and CBT among 15-year olds. Furthermore, the extent of the performance differences between girls and boys on CBT seemed to be independent of the extent of the performance differences between girls and boys on PBT (Jerrim, 2016). During the early years of computer science studies, it is frequently suggested that females' CBT performance is inferior to that of boys because they are less intensive computer users, less interested in computers and show greater anxiety in using computers than males (Cooper, 2006; Meelissen & Drent, 2008). More recent studies suggest that the gender gap is narrowing. Boys and girls perform similar on applying technical functionality. Girls seem to perform better in sharing and communication information and boys seem to score higher on self-efficacy for ICT skills (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014; Punter, Meelissen, & Glas, 2016). If students are not familiar with the digital device, it is less likely that they will perform at the same level as when they take an equivalent PBT (Davis, Janiszewska, Schwarts, & Holland, 2016). The effect of tablet familiarity on reading comprehension was investigated in a study of second-year college students (Chen, Cheng, Chang, Zheng, & Huang, 2014). The group with a high level of tablet familiarity performed significantly better than the group with a low level of tablet familiarity, suggesting that familiarity is an important consideration in digital testing.

Characteristics of the device and test items

Dadey *et al.* (2018) have argued that the variations in which test information is presented to students and the manner in which students interact with that information should be carefully considered during the process of assessment design for digital devices. Even when students are familiar with the digital testing device, its use may be more complicated in comparison with a PBT. For example, the virtual keyboard of a tablet does have limitations, which can make it more difficult to use and can result in more typing errors (Ling, 2016). Second, when looking at the posture of students when using PBT or a tablet, the investigation of Straker *et al.* (2008) found that physical discomfort such as neck and lower back pain are highly similar. Both conditions were associated with less neutral postures and greater postural and muscle activity variation than when using a desktop computer. Third, unlike in PBTs, students are often unable or less inclined to navigate between item blocks to review their answers on previous items in CBTs (Vispoel,

Hendrickson, & Bleiler, 2000). Moreover, in PBTs, students can work out problems and calculations in the margins of their booklet, while in digital mathematics tests, they have to transfer their calculations from the screen to a paper workspace. Students tend to use scrap paper to a lower extent compared to booklet margins and, therefore, make more calculation errors on CBTs than on PBTs (Johnson & Green, 2006; Kingston, 2009; Russell *et al.*, 2003).

The reading demands of an assessment may also be related to the occurrence of mode effects. Mullis, Martin, and Foy (2013) found a wide achievement gap between low reading-demand items and high reading-demand items in the TIMSS 2011 PBT in mathematics. This effect may even be stronger for digital assessments in which students read from a screen (visual fatigue) and often have to scroll (Kingston, 2009; Nardi & Ranieri, 2018). Reading on a tablet computer can take even longer than on a desktop or laptop because the screen is smaller and, therefore, requires more scrolling than with a desktop computer (Ling, 2016). In their study of Norwegian primary school students, Mangen, Walgermo, and Brønnick (2013) concluded that students reading print performed better on a reading assessment than those reading on a computer screen. They suggested that the required scrolling of the text on the computer screen was one of the reasons for the lower CBT performance. The study of Sanchez and Wiley (2009) found that scrolling negatively affects learning from screen and even could lower the capacity of the working memory. A comparison between PBTs and online assessments also indicated negative effects of scrolling on students' testing behaviour, especially among primary school students (Choi & Tinkler, 2002).

Finally, CBTs may have a negative effect on students' engagement during the assessment. In a study by Ackerman and Goldsmith (2011), the relatively poorer performance by the on-screen group compared to the PBT group was attributed to differences in the students' perception of the information presented on screen. The students seemed to regard reading from a screen more "casually" than reading from print, as they compared it with reading emails or news items. They took the test less seriously and performed worse than on the PBT. Students may also be less engaged when they are assessed with a tablet test instead of a paper or desktop computer test (Ling, 2016). A tablet offers many additional easy-to-use functions (eg, an in-built camera), which could distract students from their assessment tasks.

However, there are also advantages in the use of digital assessments, in particular, tablets, which potentially enhance students' test engagement. Students can be attracted by the use of a digital device for their assessment because of the novelty of the device or because they enjoy using the device in their daily life (Jerrim, 2016). Both computers and tablets present students with engaging, interactive items, such as interactive simulations (Fishbein, Martin, Mullis, & Foy, 2018). Furthermore, the use of tablets might be more complicated when reading on screen or when the use of scrap paper is required; at the same time, however, the touchscreen feature facilitates many actions (such as navigating), especially when students are used to touching screens (Ling, 2016).

To date, little research has been conducted on the possible positive or negative achievement effects of using tablets for summative assessments in primary education. Ling (2016) compared assessments on a tablet (iPad) and a desktop computer and found no differences in achievement scores among eighth graders. In a study of college students, Chen *et al.* (2014) investigated the effects of reading comprehension across paper, computers and tablets, and reported that the scores of the tablet group were higher than those of both the computer and paper groups, suggesting that the kind of digital device used is an important consideration in the analysis of mode effects.

Fishbein *et al.* (2018) investigated the mode effects of the eTIMSS 2019 Equivalence Study. Twenty-five countries participated in the TIMSS equivalence study. This study was conducted in the spring of 2017 to investigate whether it was necessary to adjust for mode effects in reporting

the comparisons between the CBT and PBT countries and the trends in student achievement within the CBT countries. The countries were given a choice between computers or tablets for the CBT assessment. The PBT and CBT consisted of the same items, which was a selection of the TIMSS 2015 assessment.

Fishbein *et al.* (2018) showed that with the exception of the science scores of one of the countries, the average test scores per country were lower for the CBT than for the PBT. The found mode effect could be explained by items in the CBT mode that malfunctioned during the test administration. This relates to items where students needed to draw lines on the screen and items where students needed to use the number pad. Furthermore, there were a few large items that needed scrolling. These items had substantially higher omit rates for CBT than for PBT. This indicates that the scores of the CBT countries need some adjustments in order to measure trends and compare outcomes with the PBT countries. However, the study presented no information about the extent and significance of the mode effects within each country. Moreover, no distinction was made between countries administrating the test on computers or laptops and those using tablets. In the study presented here, Dutch data for the Equivalence Study of TIMSS 2019 were used to conduct an in-depth exploration of the impact of the use of tablets in the TIMSS assessment in grade four.

Research questions

This study explores the possible differences in student achievement between the TIMSS PBT and the tablet test in relation to gender and the reading demand imposed by the mathematics items. Both gender and reading demand were mentioned in the research literature as factors that might be related to the mode effect.

The following research questions were addressed:

- What are the similarities and differences between the PBT and tablet test in terms of ability level and item parameters in the Equivalence Study in the Netherlands?
- To what extent does gender affect the possible differences between the PBT and the tablet test regarding student achievement results?
- To what extent does reading demand (high, medium or low) affect the possible differences between the PBT and the tablet test regarding student achievement results?

In line with the results of the study of Fishbein *et al.* (2018), it is expected that students will perform better on the TIMSS PBT than on the tablet test and that boys will outperform girls on the tablet test. The latter is not only consistent with the results of the literature review, but is also in line with previous TIMSS results for the Netherlands. In most of the PBT assessments since the first cycle in 1995, TIMSS has shown a (small) disadvantage for Dutch girls in mathematics and science on the PBT (Meelissen & Punter, 2016). The expectation is that this small gender difference will occur again despite the administration mode. Finally, the review of the literature suggested that differences between reading on paper and reading on screen could cause a negative mode effect for CBT and especially tablets. Additionally, that items with a high reading demand would suffer more from this negative effect compared to items with a low reading demand. Therefore, in this study, it is assumed that mathematics items with a high reading demand will show a negative mode effect for tablets.

Method

Sample

In line with the international requirements of the TIMSS Equivalence Study, the results are based on a convenience sample of Dutch primary schools. From 60 of the initially randomly

sampled Dutch primary schools, 14 schools agreed to participate in this study. The sample was enlarged by nine schools using the network of the Dutch research team. In total, 532 students from the fourth grade (10-year olds) of Dutch primary schools participated, 50% of whom were female.

Design

The TIMSS 2019 Equivalence Study employed a counterbalanced within-subjects design (Fishbein *et al.*, 2018). The students were randomly assessed twice: once on paper and once using a tablet. The administration of both tests was conducted by test administrators of the Dutch research team. The test administrators received training beforehand to ensure that the procedure for each assessment was the same and was in accordance with the international TIMSS standards. The second assessment was usually administrated 1 week after the first assessment at the same time. In the second assessment, students were presented with different items than in the first assessment. The tablets had a display of 25.5 cm and students had the possibility to hold the tablet in their hands or put the tablet down on the table during the assessment.

The test administration design was linked by common items and students, and contained 186 items from the TIMSS 2015 cycle. The items were distributed over eight different booklets. About half of the items were multiple choice, and the other half were constructed response items (Mullis & Martin, 2017). Some items needed to be adapted to make them suitable for the on-screen test (eg, "type" or "drag" instead of "write" or "draw"), but these adaptations were kept to a minimum, generally resulting in highly similar tests.

Data analysis

Item response theory (IRT) was used for a comprehensive analysis of test equivalence. IRT models pertain to a set of items constructed to form an ability scale that allows the comparison between a person's latent trait and the characteristics of an item (Embretson & Reise, 2000). The public domain software package LEXTER was used for this analysis (Glas & van Buuren, 2019).

First, it was investigated whether the convenience sample of 2017 was comparable to the original Dutch sample of 2015. For this, the raw data of the Dutch sample of 2015 and 2017 was used. This was investigated by estimating the population and item parameters, and evaluating whether differential item functioning (DIF) between the two administrations occurred. The assessment data of TIMSS 2015 were based on a representative sample of 150 schools and over 4600 Dutch grade-four students (Martin, Mullis, & Hooper, 2016; Meelissen & Punter, 2016).

Next, the PBT and CBT scores of the Equivalence Test were compared. The difference was analysed by a two-dimensional generalised partial credit model ([GPCM]; Muraki, 1992). The GPCM is commonly used in large-scale assessments for polytomously scored response items. In this study, a model including two latent variables was used to explain response behaviour: one for the PBT items and one for the CBT items (Ackerman, 1992).

The second and third research question were addressed by including gender or reading demand (only for mathematics) to the model. The categorisation of Punter, Meelissen, and Glas (2018) of the TIMSS 2015 mathematic items was used to classify the mathematics items in a low, medium or high level of reading demand.

The categorisation was based on four indicators of reading difficulty: number of words, number of different symbols, number of different specialised vocabulary words and total number of elements in the visual display (Mullis *et al.*, 2013) (See Appendix A).

Results

Equivalence study versus TIMSS 2015

Because convenience sampling was used, we checked whether this sample was a proper representation of the Dutch fourth-grade population (mean age of 10). Therefore, the raw data of the sample of the Equivalence Study was compared with the raw data of the sample of the main study of TIMSS 2015, which was administrated to over 4600 Dutch students.

The mean abilities of the eight booklets from the PBT of the Equivalence Study and the mean ability of the PBT of the TIMSS 2015 participation were compared. The comparison showed that only two out of eight booklets showed a significant medium difference. Further, the analyses showed that the test scores of the two samples were highly comparable (see Appendix B, Table B1). So, these results also indicated that the ability level of Dutch students in the Equivalence Study was comparable to the ability level of the Dutch students in the main study of TIMSS 2015.

DIF was evaluated by correlating the item parameter estimates (as rough measure of DIF) and by comparing observed and expected response frequencies (see Glas, 1999, for details of the fit statistics). The correlation between the item location parameter was high (.97 for mathematics and .96 for science) and the largest effect size of the DIF statistics was generally small (absolute differences between observed and expected item *p*-values below .03), with the exception of two items. These two items were not similar anymore in the TIMSS 2015 and TIMSS 2019 version. It turned out that the answering categories of the two items were adapted for the Equivalence Study. Therefore, these two items are omitted from this study. This finding supports the conclusion that the structure of the latent trait measured by the test was equal for the two groups and DIF was not prominent.

Paper versus tablet tests

The focus of this study was to investigate the performance differences between PBT and CBT. To investigate whether the paper and tablet tests measure similar abilities, the fit of various IRT models was evaluated using differences between the log-likelihoods. The results of this comparison are summarised in Appendix B, Table B2. This table shows that for both mathematics and science, the two-dimensional GPCM model containing eight ability distributions for the eight test versions had the best model fit.

However, based on the two-dimensional GPCM model (with eight ability distributions, one for each booklet), the ability scales of the two tests were found to be highly correlated: mathematics .91 and science .88. Therefore, we concluded that there is reasonable support for the hypothesis that the paper and tablet tests seem to measure the same abilities. A subsequent analysis was to investigate the extent to which the booklets differed within a test mode and between test modes (paper vs. tablet). This was first done for the mathematics domain (see Appendix B, Table B3). Within each test mode, no booklet had a *z*-score for differences in mean ability above or below the critical Z-values of −1.96 and 1.96 for mathematics. This means that no significant differences were found within either the PBT or CBT mode.

Similar analyses were performed for the comparison between the test modes for the homogeneous booklets in the mathematics domain. No booklet between either of the test modes demonstrated a *z*-score for differences in mean ability above or below the critical Z-values of −1.96 and 1.96. This means that no significant differences were found between the test modes based on the homogeneous booklets (Table 1).

These analyses were subsequently conducted for the science domain. The same results for the mathematics domain were found within the science domain. There were no significant differences

between the booklets within either test mode, and no booklet demonstrated a *z*-score above or below the critical Z-values of −1.96 and 1.96. Thus, no significant differences were found within the PBT and CBT modes for the science domain (see Appendix B, Table B4).

For the comparison between the test modes for the homogenous booklets in the science domain, the same results were found between each test mode, and no booklet had a *z*-score for differences in mean ability above or below the critical Z-values of −1.96 and 1.96. Thus, no significant differences were found between the test modes based on the homogenous booklets in the science domain (Table 2).

As above, DIF was evaluated by correlating the item parameter estimates and by comparing observed and expected response frequencies. The correlation was high (.95 for mathematics and .92 for science) and effect sizes were small (absolute differences between observed and expected item *p*-values below .03).

Student characteristics: girls versus boys

In general, the girls scored significantly higher on the TIMSS Equivalence Test (mathematics and science (Z-value = 2.25; $p < .05$). However, the effect size (Cohen's $d = 0.13$) can be considered small (Field, 2009).

Table 1: Differences between mean abilities of the booklets in the mathematics domain between the PBT and tablet tests

	Δ (μ paper − μ tablet)	Δ (SE paper − SE tablet)	Z-value
Booklet_1	−0.21	0.35	−0.60
Booklet_2	−0.22	0.40	−0.55
Booklet_3	−0.37	0.46	−0.81
Booklet_4	−0.12	0.50	−0.23
Booklet_5	−0.15	0.41	−0.37
Booklet_6	−0.51	0.50	−1.02
Booklet_7	0.08	0.37	0.22
Booklet_8	–	–	–

Note: Booklet_8 was used as the reference group to identify the latent scales for paper and tablet assessment. Booklet scores that differed significantly from the reference group are indicated by * (alpha < .05) or ** (alpha < .01).

Table 2: Differences between mean abilities of the booklets in the science domain between the PBT and tablet tests

	Δ (μ paper − μ tablet)	Δ (SE paper − SE tablet)	Z-value
Booklet_1	−0.11	0.37	−0.60
Booklet_2	0.24	0.39	−0.55
Booklet_3	0.40	0.44	−0.81
Booklet_4	0.30	0.47	−0.23
Booklet_5	0.17	0.48	−0.37
Booklet_6	0.37	0.39	−1.02
Booklet_7	0.08	0.38	0.22
Booklet_8	–	–	–

Note: Booklet_8 was used as the reference group to identify the latent scales for paper and tablet assessment. Booklet scores that differed significantly from the reference group are indicated with * (alpha < .05) or ** (alpha < .01).

Subsequently, the performance of both boys and girls on the test was compared in combination with the test mode (interaction effect). The achievement results on the PBT did not differ between boys and girls. However, a significant difference was found between boys and girls regarding their performance on the tablet test. Girls scored higher in the CBT mode (*Z*-value = 2.3; $p < .05$). The effect size was small (Cohen's $d = 0.15$).

Finally, the differences between the reading demand categories, both within and between the test modes, were examined. The mathematics items were categorised in three levels: low, medium or high reading demand (Mullis *et al.*, 2013; Punter *et al.*, 2018). The categorisation of the items can be found in Appendix A. For this categorisation, a two-dimensional GPCM model with one marginal was analysed. The descriptive statistics of the item location parameter, which are based on the reading demand categories, are presented in Table 3. Higher means indicate more difficult items, though the differences in Table 3 are not significant.

First, when controlling for reading demand, the correlation between the two dimensions remained high ($r = .91$), indicating that the response behaviour of the students on the test is best explained by two latent abilities. The results showed no significant differences within the test modes or between the levels of reading demand. This means that the students' performance was not related to the reading demand of the items in neither of the test modes. The results for the paper mode are shown in Appendix B, Table B5 and for the tablet mode in Appendix B, Table B6.

Finally, the differences between the PBT and CBT, as categorised by the reading demand, were calculated, as shown in Table 4. This suggests that the level of reading demand was not related to the item difficulty experienced or the students' performance on either the PBT or the CBT.

Limitations of this study

One of the limitations of the present study is the convenience sample of 23 schools, which is not representative of all grade-four students in the Netherlands. A comparison between the Dutch results of TIMSS 2015 and the PBT of the Equivalence Study demonstrated no differences in

Table 3: Descriptive statistics of the item location parameter based on the reading demand categories and test mode

Reading demand category	*Test mode*	*Number of items*	*Mean*	*SE*
Low	Paper	30	−0.845	0.688
	Tablet	30	−0.474	0.797
Medium	Paper	35	−0.054	0.596
	Tablet	35	0.862	1.026
High	Paper	23	0.139	0.547
	Tablet	23	0.145	0.369

Note: Test characteristics: low versus high reading demand in the mathematics items

Table 4: Differences between the paper and tablet modes within the reading demand category

Reading Demand Category	*Δ (μ paper − μ tablet)*	*Δ (SE paper − SE tablet)*	*Z-value*
Low	−1.319	1.035	1.253
Medium	1.187	1.187	−0.772
High	−0.006	0.660	−0.009

students' ability level and the functioning of the test items in the mathematics and science assessments. However, other background characteristics of the students participating in the current study may still have had an impact on the results. For example, it might be that primary schools which are using tablets more frequently for instruction and assessments were more interested in participating in this study than other schools. Therefore, at these schools, the students' high familiarity with tablets might be related to the absence of a mode effect in this study.

Second, the short eTIMSS questionnaire at the end of the CBT focused on digital devices in general and not specially on the device used for the assessment. Therefore, it was not possible to make a distinction between "heavy" and "light" users of tablets to reveal how this might have affected performance. Furthermore, almost all Dutch grade-four students in this study responded that they had access to both computers and tablets, both at home and at school. Collecting this information in future TIMSS studies would provide in-depth information about the advantages and disadvantages of tablet assessments in relation to achievement and could also be used to improve the criteria for the development of new items specifically designed for CBTs.

Conclusion and discussion

This study investigated the occurrence of mode effects among Dutch grade-four students in the Equivalence Study of TIMSS 2019. The results suggest that it does not seem to matter whether Dutch grade-four students are presented with mathematics and science problems on paper or tablet. The two test modes produced similar measurements for both mathematics and science, since there were no significant differences between the student ability scales and the item parameters. However, what should be taken in consideration is that girls performed marginally better on the tablet test than boys. This small difference should be taken into account for the TIMSS 2019 main study, to see whether this difference lasts. This could indicate a test mode effect between subgroups.

The overall result seems to contradict the conclusions of Fishbein *et al*. (2018), who found that, on average, students from the participating TIMSS countries performed better on the paper version of the Equivalence Study. The lack of significant mode effects among Dutch grade-four students may be explained by their high level of familiarity with tablet devices. This can be supported with the data of the Dutch Youth Institute in 2015, where it was estimated that children owned one or more tablets in well over 75% of Dutch households (Nederlands Jeugd Instituut, 2015). Therefore, it is very likely that the majority of Dutch grade-four students have regular access to tablets, whether it is in school, home or both. Familiarity with a device and its features (eg, touch screen) is regarded as a key consideration concerning comparability with paper tests (Davis *et al.*, 2016).

The comparison between the results of this study of Dutch data and of the study of Fishbein *et al*. (2018) at the country level suggests the occurrence of a mode effect differs between countries. To monitor the switch of the test mode, countries participating in the digital assessment in the main study of TIMSS 2019 also had to administer the TIMSS assessment on paper in an additional number of schools. The results of the so-called Bridge Study are important in order to correct for the varying mode effects between countries. The corrections will be done via a linear transformation that adjusts for the mode effect for each country in the same way. Since there is no mode effect found for the Netherlands, it is difficult to predict how this overall correction will affect the Dutch results of TIMSS. Additionally, it is uncertain what this will mean for the trend analysis within the Netherlands and the ranking between countries.

Although this study did not find a mode effect, it does not mean that some of the advantages and disadvantages of CBT described in the literature did not occur during the tablet assessment.

A study of university students revealed that although students preferred the CBT and generally performed better on it, they also experienced limitations, mainly because the test items were presented on a screen (Nardi & Ranieri, 2018). During the data collection for the present study, some of the advantages and disadvantages of the tablet assessment were reported by the Dutch test administrators. They observed that students faced functional difficulties when performing eTIMSS. Some features on the tablet were not completely functional. Thereby students who were assigned to the PBT for the second time seemed less motivated by it than students who were assigned to the tablet test for the second time. However, information about the students' experiences of using the tablet test or their preferences for a certain test mode was not systematically collected. Therefore, it was not possible to systematically check the students' preferences and motivation.

The study also examined the occurrence of a mode effect between mathematics items with a high, medium or low reading demand (Mullis *et al.*, 2013; Punter *et al.*, 2018). The literature review suggested that the complexity and length of reading texts on a computer or tablet screen were two of the main disadvantages of CBT and could be related to the lower performance of students on CBTs compared to PBTs (Choi & Tinkler, 2002; Kingston, 2009; Ling, 2016; Mangen *et al.*, 2013; Nardi & Ranieri, 2018). Therefore, a negative mode effect for the tablet test was expected for mathematics items classified as having a high reading demand. Regardless of the test mode used, Dutch students performed equally on these items. Perhaps this disadvantage occurs specifically in cases of tablet testing, as students have to scroll back and forth in order to give an answer. In the tablet version of the TIMSS test, the scrolling within the test items with a high reading demand was kept to a minimum.

The present study found that girls slightly outperformed boys on the tablet test, and that boys and girls performed equally on the PBT. The Dutch results of earlier TIMSS PBTs showed that boys usually scored significantly higher in mathematics and science than girls (Meelissen & Punter, 2016). However, the extent of these differences has fluctuated over the assessment years. In the most recent TIMSS cycle of 2015, the Dutch gender gap favouring boys in mathematics was significant but very small, and there was no significant gap in science achievement. This may explain why no significant gender differences were found this time around on the PBT. The current study revealed that girls slightly outperformed boys on the tablet assessment. This confirms the findings of previous studies indicating that the traditional gender gap in computer skills is changing (Fraillon *et al.*, 2014; Punter *et al.*, 2016). However, further investigation is needed regarding the specific skills, efficacy and attitudes of girls and boys towards tablet assessments, making this a relevant subject for future studies (Dündar & Akçayir, 2014; Pruet, Ang, & Farzin, 2016).

A major advantage of digital assessments is that information about student testing behaviour (such as response time) is often available. A study by Soland (2018) showed that boys are more likely to show rapid-guessing behaviour (responding without taking sufficient time to understand the item) compared to girls. Boys may be more inclined to skip and click to the next item than girls, especially if it concerns a low-stakes assessment, such as TIMSS. The Equivalence Study did not offer this kind of response data. In the future, it would be relevant to examine possible gender differences in testing behaviour and to explore whether these differences affect the mathematics and science achievement of girls and boys.

Finally, to be able to compare the PBT with the CBT, the assessment of the Equivalence Study consisted of trend items from the TIMSS 2015 PBT. This means that the students were presented with thoroughly tested items as well as with items which were not developed with a CBT in mind. One of the advantages of the CBT is that students can be presented with engaging interactive items, such as simulated science investigations. For the main data collection, TIMSS has developed so-called problem-solving inquiring (PSI) items. A PSI is presented as a small story with a

number of related mathematics or science test items. Part of these items takes the form of interactive simulations. It would be relevant to explore whether the achievement of students on PSI items differs from that on comparable "traditional" items in terms of the requisite knowledge and skills described in the TIMSS assessment framework and whether these PSIs influence gender differences in mathematics and science achievement.

Acknowledgements
This work was supported by the Netherlands Initiative for Educational Research (NRO) (Grant number: 405-17-321).

Statements on open data, ethics and conflict of interest
The authors would like to state that there is no potential conflict of interest in this study. They would also like to declare that the work has not been published previously.

In collaboration with the IEA (International Association for the Evaluation of Educational Achievement) and the International TIMSS and PIRLS Study Centre, the University of Twente followed the strictest guidelines to coordinate and administer the assessment. The personal information of the subjects in this study is not available. In addition, the subjects were informed that participation in the test was voluntary and would not affect their grades.

The experimental data can be provided upon request.

References
Ackerman, T. A. (1992). Didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91.

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied*, *17*(1), 18–32. https://doi.org/10.1037/a0022086

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. *The Journal of Technology, Learning, and Assessment*, *6*(9), 1–39.

Brummelhuis, A., & Binda, A. (2017). *Vier in balans-monitor 2017: De hoofdlijn* [Four in balance monitor 2017: Main results]. Retrieved from https://www.kennisnet.nl/fileadmin/kennisnet/publicatie/vierinbalans/Vier-in-balans-monitor-2017-Kennisnet.pdf

Chen, G., Cheng, W., Chang, T. W., Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter? *Journal of Computers in Education*, *1*(2), 213–225. https://doi.org/10.1007/s40692-014-0012-z

Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper and computer-based assessment in a K-12 setting*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans. Retrieved from https://www.researchgate.net/profile/Seung_Choi2/publication/274713232_Evaluating_comparability_of_paper-and-pencil_and_computer-based_assessment_in_a_K-12_setting_1/links/55275ead0cf2e486ae40feb8.pdf

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.

Cooper, J. (2006). The digital divide: The special case of gender. *Journal of Computer Assisted Learning*, *22*, 320–334. https://doi.org/10.1111/j.1365-2729.2006.00185.x

Dadey, N., Lyons, S., & DePascale, C. (2018). The comparability of scores from different digital devices: A literature review and synthesis with recommendations for practice. *Applied Measurement in Education*, *31*(1), 30–50. https://doi.org/10.1080/08957347.2017.1391262

Davis, L. L., Janiszewska, I., Schwarts, R., & Holland, L. (2016). *NAPLAN device effects study*. Retrieved from https://nap.edu.au/docs/default-source/default-document-library/naplan-online-device-effect-study.pdf?sfvrsn=2

Dündar, H., & Akçayir, M. (2014). Implementing tablet PCs in schools: Students' attitudes and opinions. *Computers in Human Behavior*, *32*, 40–46. https://doi.org/10.1016/j.chb.2013.11.020

Embretson, S., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Earlbaum.

Faber, J. M., & Visscher, A. J. (2016). *De effecten van Snappet. Effecten van een adaptief onderwijsplatform op leerresultaten en motivatie van leerlingen* [The effects of Snappet. Effects of an adaptive learning platform on the achievement and motivation of students]. Retrieved from https://www.kennisnet.nl/fileadmin/kennisnet/leren_ict/leren_op_maat/bijlagen/De_effecten_van_Snappet_Universiteit_Twente.pdf

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Los Angeles, CA: Sage.

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education*, *6*(11), 1–23. https://doi.org/10.1186/s40536-018-0064-z

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age. The IEA International and Information Literacy Study international Report*. Retrieved from https://research.acer.edu.au/cgi/viewcontent.cgi?article=1009&context=ict_literacy

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. *Journal of Educational Measurement*, *39*(2), 133–147. https://doi.org/10.1111/j.1745-3984.2002.tb01139.x

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Pscyhometrika*, *64*, 273–294.

Glas, C. A. W., & van Buuren, N. (2019). LEXTER. [Public domain computer software]. Retrieved from https://shinylexter.com/

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour and Information Technology*, *33*(4), 410–422. https://doi.org/10.1080/0144929X.2012.710647

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

Johnson, M., & Green, S. (2006). On-line mathematics: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning and Assessment*, *4*(5), 4–35.

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education*, *22*(1), 22–37. https://doi.org/10.1080/08957340802558326

Ling, G. (2016). Does it matter whether one takes a test on an iPad or a desktop computer? *International Journal of Testing*, *16*, 352–377. https://doi.org/10.1080/15305058.2016.1160097

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, *58*, 61–68. https://doi.org/10.1016/j.ijer.2012.12.002

Martin, M. O., & Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *TIMSS: Methods and procedures in TIMSS 2015*. Retrieved from http://timssandpirls.bc.edu/publications/timss/2015-methods.html

Meelissen, M. R. M., & Drent, M. (2008). Gender differences in computer attitudes: Does the school matter? *Computers in Human Behavior*, *24*, 969–985. https://doi.org/10.1016/j.chb.2007.03.001

Meelissen, M. R. M., & Punter, R. A. (2016). *Twintig jaar TIMSS: Ontwikkelingen in leerlingprestaties in de exacte vakken in het basisonderwijs, 1995–2015* [Twenty years of TIMSS: Developments in student achievement in mathematics and science in primary education, 1995–2015]. Retrieved from https://ris.utwente.nl/ws/portalfiles/portal/5136442

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 assessment frameworks*. Retrieved from https://timssandpirls.bc.edu/timss2019/frameworks/

Mullis, I. V. S., Martin, M. O., & Foy, P. (2013). *TIMSS and PRILS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade-implications for early reading*. Retrieved from https://timssandpirls.bc.edu/timsspirls2011/international-database.html

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Nardi, A., & Ranieri, M. (2018). Comparing paper-based and electronic multiple-choice examinations with personal devices: Impact on students' performance, self-efficacy and satisfaction. *British Journal of Educational Technology*, *50*(3), 1495–1506. https://doi.org/10.1111/bjet.12644

Nederlands Jeugd Instituut. (2015). *Factsheet media in het gezin [Fact sheet: The use of media in families]*. Retrieved from https://www.nji.nl/nl/Download-NJi/Publicatie-NJi/Factsheet-Media-in-het-gezin.pdf

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics*, *51*(9), 1352–1375.

Poll, H. (2015). *Pearson student mobile device survey 2015*. Retrieved from http://www.pearsoned.com/wp-content/uploads/2015-Pearson-Student-Mobile-Device-Survey-College.pdf

Pruet, P., Ang, C. S., & Farzin, D. (2016). Understanding tablet computer usage among primary school students in underdeveloped areas: Students' technology experience, learning styles and attitudes. *Computers in Human Behavior*, 55, 1131–1144. https://doi.org/10.1016/j.chb.2014.09.063

Punter, R. A., Meelissen, M. R. M., & Glas, C. A. W. (2016). Gender differences in computer and information literacy: An exploration of the performances of girls and boys in ICILS 2013. *European Educational Research Journal*, *16*(6), 762–780. https://doi.org/10.1177/1474904116672468

Punter, R. A., Meelissen, M. R. M., & Glas, C. A. W. (2018). An IRT model for the interaction between item properties and group membership: Reading demand in the TIMSS-2015 mathematics test. In R. A. Punter (Ed.), *Improving the modelling of response variation in international large-scale assessments* (Published doctoral dissertation). Enschede, the Netherlands: University of Twente. https://doi.org/10.3990/1.9789036546867

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: A look back into the future. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 279–293. https://doi.org/10.1080/0969594032000148145

Sanchez, C., & Wiley, J. (2009). To scroll or not to scroll: Scrolling, working memory capacity, and comprehending complex texts. *Human Factors*, *51*(5), 730–738. https://doi.org/10.1177/0018720809352788

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, *33*(1), 1–19. https://doi.org/10.1111/jcal.12172

Soland, J. (2018). The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. *Applied Measurement in Education*, *31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

Straker, L., Coleman, J., Skoss, R., Maslen, N. A., Burgess-Limerick, R., & Pollock, C. M. (2008). A comparison of posture and muscle activity during tablet computer, desktop computer and paper use by young children. *Ergonomics*, *51*(4), 540–555.

Vispoel, W., Hendrickson, A., & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement*, *37*(1), 21–38.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2007). A meta-analysis of testing mode effects in Grade K-12 mathematics tests. *Educational and Psychological Measurement*, *67*, 219–238.

APPENDIX A

Table A1: Categorisation of the mathematics items based on reading demand levels low (1), medium (2) or high (3)

			Technical words		Symbolic language		Visual display			
Item	*Reading demand level*	*Words*	*Unique*	*Total*	*Unique*	*Total*	*Density & interaction*	*Density only*	*Interaction only*	*Items deleted in 2017 study*
M061275	1	0	0	0	9	10	0	0	0	
M061027	2	30	3	3	5	5	0	0	0	
M061255	3	62	1	3	5	5	0	0	0	
M061021	2	18	2	3	2	3	3	2	1	
M061043	2	21	0	0	2	2	6	5	1	
M061151	2	21	0	0	5	23	0	0	0	
M061172	1	10	1	1	11	12	0	0	0	
M061223	2	17	0	0	12	17	4	3	1	
M061269	1	11	0	0	0	0	12	6	6	
M061081A	2	8	2	2	5	6	5	4	1	
M061081B	2	8	2	2	5	6	5	4	1	
M061026	1	8	4	4	8	9	0	0	0	
M061273	1	0	0	0	8	8	0	0	0	
M061034	2	34	0	0	3	3	0	0	0	
M061040	2	25	3	5	5	6	3	3	0	
M061228	3	37	1	2	1	1	0	0	0	
M061166	1	9	1	1	5	9	0	0	0	
M061171	3	31	0	0	6	30	0	0	0	
M061080	2	15	1	2	1	1	2	1	1	
M061222	2	15	0	0	12	17	0	5	3	
M061076	3	31	1	3	1	1	13	11	2	Deleted item
M061084	3	34	1	1	3	3	13	11	2	Deleted item
M051206	1	0	0	0	4	4	0	0	0	
M051052	1	9	0	0	5	5	0	0	0	
M051049	2	21	1	1	5	5	0	0	0	
M051045	2	26	1	1	3	4	0	0	0	
M051098	1	5	2	2	5	5	0	0	0	
M051030	2	25	0	0	2	2	0	0	0	
M051502	3	81	1	1	2	5	18	9	9	
M051224	1	10	1	1	0	0	8	4	4	

Table A1: Continued

Item	Reading demand level	Words	Technical words		Symbolic language		Visual display			Items deleted in 2017 study
			Unique	Total	Unique	Total	Density & interaction	Density only	Interaction only	
M051207	2	12	2	2	0	0	12	8	4	
M051427	2	17	2	3	13	20	4	3	1	
M051533	3	24	1	1	5	12	10	8	2	
M051080	3	49	0	0	6	9	14	12	2	Deleted item
M061018	1	17	2	2	4	4	0	0	0	
M061274	1	0	0	0	8	8	0	0	0	
M061248	3	57	1	1	3	3	0	0	0	
M061039	2	24	1	1	1	1	0	0	0	
M061079	2	24	2	2	0	0	6	4	2	
M061179	1	13	2	2	10	11	0	0	0	
M061052	3	38	1	3	5	25	5	5	0	
M061207	2	7	2	2	12	14	3	2	1	
M061236	2	16	2	3	4	8	6	5	1	
M061266	3	44	3	6	2	2	20	17	3	
M061106	3	33	0	0	4	4	20	16	4	
M051401	2	21	0	0	5	6	0	0	0	
M051075	1	5	2	2	5	5	0	0	0	
M051402	1	16	1	2	2	2	0	0	0	
M051226	2	8	0	0	1	1	12	8	4	
M051131	1	12	1	1	4	7	0	0	0	
M051103	1	0	0	0	8	8	0	0	0	
M051217	1	8	0	0	6	7	3	2	1	
M051079	2	19	4	4	6	11	7	6	1	
M051211	2	19	1	1	0	0	38	33	5	
M051102	1	16	3	4	7	8	0	0	0	
M051009	2	25	1	1	11	11	8	7	1	
M051100	3	58	0	0	4	4	28	23	5	
M061178	1	12	1	1	3	3	0	0	0	
M061246	1	8	2	2	5	5	0	0	0	
M061271	1	0	0	0	4	4	0	0	0	
M061256	3	64	1	2	9	14	9	8	1	
M061182	2	24	1	3	2	2	0	0	0	

Table A1: Continued

Item	Reading demand level	Words	Technical words		Symbolic language		Visual display			Items deleted in 2017 study
			Unique	Total	Unique	Total	Density & interaction	Density only	Interaction only	
M061049	1	13	2	2	10	11	0	0	0	
M061232	2	21	1	1	7	23	0	0	0	
M061095	3	34	2	6	5	11	33	31	2	
M061264	3	79	0	0	6	12	6	4	2	
M061108	2	22	0	0	0	0	6	4	2	
M061211A	3	36	0	0	1	3	36	31	5	
M061211B	3	52	0	0	5	7	11	7	4	
M051043	1	11	1	1	8	9	9	8	1	
M051040	2	9	0	0	1	1	22	18	4	
M051008	3	54	1	2	4	4	2	2	0	
M051031A	3	39	2	2	1	1	5	4	1	
M051031B	3	39	2	2	1	1	5	4	1	
M051508	3	64	2	5	14	19	0	0	0	
M051216A	2	32	1	1	0	0	7	6	1	
M051216B	2	32	1	1	0	0	7	6	1	
M051221	1	9	1	1	4	4	2	1	1	
M051115	1	12	0	0	0	0	24	20	4	
M051507A	3	43	0	0	13	25	11	10	1	
M051507B	3	46	0	0	13	26	11	10	1	
M061240	2	22	3	8	9	19	0	0	0	
M061254	3	63	2	2	9	10	21	20	1	Deleted item
M061244	2	31	0	0	4	8	0	0	0	
M061041	1	5	0	0	6	6	0	0	0	
M061173	1	10	1	1	5	8	0	0	0	
M061252	2	31	0	0	6	14	0	0	0	
M061261	3	47	1	1	8	12	0	0	0	
M061224	2	17	3	4	8	8	6	5	1	
M061077	1	8	0	0	0	0	23	18	5	
M061069A	3	33	1	4	12	17	9	8	1	
M061069B	3	41	1	3	12	17	9	8	1	

APPENDIX B

Table B1: Comparison of ability estimates between the TIMSS 2015 main study sample and the equivalence study sample

	Mean	*SD*	*SE (Mean)*	*Z-value*	*Effect size*
Norm group 2015	0.00	1.00	0.00		
Booklet 1	−0.22	0.84	0.11	−2.02*	−0.26
Booklet 2	−0.21	0.87	0.12	−1.78	
Booklet 3	−0.04	0.90	0.12	−0.31	
Booklet 4	−0.28	1.07	0.13	−2.08*	−0.26
Booklet 5	0.11	1.03	0.13	0.09	
Booklet 6	0.02	0.95	0.12	0.02	
Booklet 7	0.03	1.03	0.13	0.03	
Booklet 8	−0.05	1.31	0.17	−0.3	

Note: Booklet scores showing a significant difference, alpha < .05, to the norm group of 2015 are indicated by *. 0.20 < absolute effect size < 0.30 is a medium effect size.

Table B2: Comparison of IRT models based on model fit for mathematics and science

	Δ *df*		Δ*-2ll*		*p*	
	M	*S*	*M*	*S*	*M*	*S*
1-dim PCM vs. 1-dim GPCM	187	215	360	408	<0.001	<0.001
1-dim GPCM vs. 2-dim GPCM	1	1	12	18	<0.001	<0.001
2-dim GPCM vs. 2-dim GPCM (8 distributions)	35	35	381	206	<0.001	<0.001

Note: "M" symbolises mathematics, and "S" symbolises science.

Table B3: Differences between mean abilities of the booklets in the mathematics domain within the PBT and the tablet mode

	Mean		*SE*		*Z-value*	
	Paper	*Tablet*	*Paper*	*Tablet*	*Paper*	*Tablet*
Booklet_1	−0.10	0.10	0.29	0.21	−0.38	0.51
Booklet_2	−0.10	0.13	0.30	0.27	−0.03	0.46
Booklet_3	0.04	0.41	0.31	0.34	0.13	1.23
Booklet_4	−0.34	−0.23	0.32	0.38	−1.05	−0.60
Booklet_5	0.11	0.27	0.30	0.29	0.38	0.92
Booklet_6	−0.08	0.43	0.27	0.43	−0.30	1.00
Booklet_7	0.10	0.03	0.26	0.27	0.40	0.10
Booklet_8	0	0	1.00	1.00		

Note: Booklet_8 was used as the reference group. Booklet scores that differed significantly from the reference group are indicated by * (alpha < .05) or ** (alpha < .01).

Table B4: Differences between mean abilities of the booklets in the science domain within the PBT and the tablet mode

	Mean		*SE*		*Z-value*	
	Paper	*Tablet*	*Paper*	*Tablet*	*Paper*	*Tablet*
Booklet_1	−0.25	−0.14	0.24	0.28	−1.06	−0.51
Booklet_2	−0.003	−0.24	0.26	0.30	−0.01	−0.81
Booklet_3	0.10	−0.30	0.28	0.34	0.36	−0.87
Booklet_4	0.11	−0.19	0.28	0.35	0.40	−0.55
Booklet_5	0.24	−0.07	0.31	0.37	0.77	−0.18
Booklet_6	0.29	−0.07	0.28	0.27	1.04	−0.26
Booklet_7	0.006	−0.08	0.27	0.27	1.04	−0.28
Booklet_8	0	0	1.00	1.00		

Note: Booklet_8 was used as the reference group. Booklet scores that differed significantly from the reference group are indicated by * (alpha < .05) or ** (alpha < .01).

Table B5: Differences between levels of reading demand within the paper mode

Reading demand	Δ *Mean*	Δ *SE*	*Z-value*
Low–medium	−0.791	0.910	−0.869
Medium–high	−0.193	0.809	−0.239
Low–high	−0.879	0.879	−1.12

Table B6: Differences between levels of reading demand within the tablet mode

Reading demand	Δ *Mean*	Δ *SE*	*Z-value*
Low–medium	−1.336	1.300	−1.028
Medium–high	0.717	1.09	0.658
Low–high	−0.619	0.878	−0.705