



# Effects of multimedia on psychometric characteristics of cognitive tests: A comparison between technology-based and paper-based modalities

De Van Vo <sup>a,b,c,\*</sup>, Benő Csapó <sup>d,e,\*\*</sup>

<sup>a</sup> Faculty of Education, An Giang University- VNU-HCM, 18 Ung Van Khiem St, Dong Xuyen Ward, Long Xuyen City, An Giang Province, Vietnam

<sup>b</sup> Doctoral School of Education, University of Szeged, 32-34. Petőfi S. sgt., Szeged H-6722, Hungary

<sup>c</sup> Epi•STEM National Centre for STEM Education, School of Education, University of Limerick, Plassey Park Road, Casteltroy, Limerick, Ireland

<sup>d</sup> Institute of Education, University of Szeged, 32-34. Petőfi S. sgt., Szeged H-6722, Hungary

<sup>e</sup> MTA-SZTE Research Group on the Development of Competencies, Szeged, Hungary

## ARTICLE INFO

### Keywords:

Test equivalence

DBF

DIF

Inductive reasoning

Control-of-variables strategy

Technology-based assessment

Paper-and-pencil test

## ABSTRACT

The study aims to investigate the effects of delivery modalities on psychometric characteristics and student performance on cognitive tests. A first study assessed the inductive reasoning ability of 715 students under the supervision of teachers. A second study examined 731 students' performance on the application of the control-of-variables strategy in basic physics but without teacher supervision due to the COVID-19 pandemic. Rasch measurement showed that the online format fitted to the data better in the unidimensional model across two conditions. Under teacher supervision, paper-based testing was better than online testing in terms of reliability and total scores, but contradictory findings were found in turn without teacher supervision. Although measurement invariance was confirmed between two versions at item level, the differential bundle functioning analysis supported the online groups on the item bundles constructed of figure-related materials. Response time was also discussed as an advantage of technology-based assessment for test development.

## 1. Introduction

Information and Communications Technology (ICT) has become increasingly relevant in most aspects of modern life, including work and school. Computers have been important supplementary tools in the teaching and learning process. As part of this explosion of new technologies, technology-based assessment (TBA) is being used globally on either computers or other electronic devices, such as smartphones, tablets and other portable devices. TBA provides evidence of positive results on student learning performance, motivation and attitudes (Mohamadi, 2018; Nikou & Economides, 2018; Sheard & Chambers, 2014). TBA offers numerous benefits, such as a higher standardization of test administration, efficient test scoring, and the likelihood of immediate reporting and interpreting of results (Csapó, Ainley, Bennett, Latour, & Law, 2012; Shute & Rahimi, 2017). A large number of institutions have considered introducing TBA administration in recent

years, and thus TBA has progressively replaced traditional paper-and-pencil assessment.

Creativity, critical thinking, problem-solving and ICT are all regarded as significant aptitudes in the 21st century (Voogt & Roblin, 2012). Moreover, students are expected to demonstrate these abilities and skills to meet the demands of future jobs. Inductive reasoning and scientific reasoning in the control-of-variables strategy are closely tied to problem-solving and play an important role in learning core school disciplines (e.g., Adey & Csapó, 2012; Chen & Klahr, 1999; Van Vo & Csapó, 2021a). Modern society is under ever greater pressure to deal with more information in a shorter amount of time. Assessment of these skills has therefore come under increasing consideration. Like other testing in educational contexts, these cognitive tests can be used in testing mechanisms in a virtual environment, so they have gradually transformed from traditional forms into technology-rich formats.

The paper method is restricted to using static text and graphics,

*Abbreviations:* DBF, differential bundle functioning; DIF, differential item functioning.

\* Correspondence to: Epi•STEM National Centre for STEM Education, School of Education, University of Limerick, Limerick, Ireland.

\*\* Corresponding author at: Institute of Education, University of Szeged, 32-34. Petőfi S. sgt., Szeged H-6722, Hungary.

E-mail addresses: [vvde@agu.edu.vn](mailto:vvde@agu.edu.vn) (D. Van Vo), [csapo@edpsy.u-szeged.hu](mailto:csapo@edpsy.u-szeged.hu) (B. Csapó).

<sup>1</sup> ORCID: <https://orcid.org/0000-0002-8515-0221>

<sup>2</sup> ORCID: <https://orcid.org/0000-0001-7550-6354>

<https://doi.org/10.1016/j.stueduc.2023.101254>

Received 14 May 2021; Received in revised form 26 May 2022; Accepted 26 February 2023

Available online 2 March 2023

0191-491X/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

whereas computers are capable of presenting rich visualizations of figures and even dynamic interactions with test-takers (Greiff, Molnár, Martin, Zimmermann, & Csapó, 2018). However, previous studies have reported inconsistent findings when comparing performance between dual modes of administration (Gates & Kochan, 2015; Williamson, Williamson, & Hinze, 2017). There are various approaches to investigating equivalent groups in these studies, such as construct validity and mean scores from test to item levels (Gates & Kochan, 2015; Williamson et al., 2017).

In the context of Vietnam, most schools were at the *infusing* stages of ICT-facilitated teaching and learning pedagogies, and the conditions in schools were suitable for the necessary transformation of their ICT-facilitated teaching and learning practices (Maftuh, 2011). Nevertheless, the development of assessment was thought to be at the *applying* stage, meaning that teachers had started learning to use ICT along with traditional methods. In recent years, schools have been interested in using technology in educational assessment, although paper-and-pencil mode is still used officially in Vietnam. Institutions often offer both paper and TBA administration because of the inadequacies of the current ICT infrastructure. Specifically, when novel coronavirus (COVID-19) broke out, it posed a major challenge for the traditional school environments. Testing administration in education should be in line with the national restrictions on social interactions due to the pandemic. Therefore, studies comparing the impact of the paper-based and TBA modes of administration warrant a scholarly inquiry. The main aim of our study is to investigate the effects of technology in different modalities in cases with and without supervision on cognitive tests.

## 2. Inductive reasoning and control-of-variables strategy

Inductive reasoning (IR) is a cognitive process in which one is expected to draw a general conclusion based on particular facts or individual cases (Adey & Csapó, 2012; Sternberg & Sternberg, 2012). It has been regarded as one of the seven primary mental abilities to contribute to intelligent behavior (Kinshuk & McNab, 2006). IR plays a more important role when complex or unfamiliar problems occur, for which no specific content knowledge is applicable; based on observation, inductive processes can be applied to generate hypothetical rules, and these relational systems in the problems can be modeled (Perret, 2015). Several kinds of tasks have been suggested in the literature as ways of measuring inductive reasoning proficiency. The most popular tasks are verbal and geometric analogies, number series completions, classifications and geometric matrices (Adey, Csapó, Demetriou, Hautamäki, & Shayer, 2007; Sternberg & Sternberg, 2012; Van Vo & Csapó, 2022). In this paper, we refer the term task (or subtest) as a bundle of items which are classified based on a typical structure or a target subskill in cognitive tests.

Control-of-variables strategy (CVS) in scientific reasoning mentions to the scientific process skills that enable students to construct their arguments and to argue on the basis of evidence (Schwichow, Christoph, Boone, & Härtig, 2016). A CVS item consists of a complex reasoning pattern or strategy within several reasoning schemes to examine whether or not an experimental system can conclude a reasonable result if certain variable conditions change (Adey & Csapó, 2012; Chen & Klahr, 1999). CVS is an acquisitive contribution to the development of scientific reasoning skills because it involves inquiring about the components of experimental sets (Chen & Klahr, 1999). Therefore, school curricula have considered developing this skill for students at various education levels (Wood, Koenig, & Owens, 2018). Chen and Klahr (1999) have classified CVS into four kinds of tasks to assess four sub-skills accordingly: *planning* controlled experiments, *identifying* controlled experiments, *interpreting* controlled experiments and *understanding* the indeterminacy of confounded experiments.

## 3. Studies on comparisons of modes of administration

TBA has been highly beneficial for students, teachers, test administrators and other stakeholders. TBA supports the collection of reliable data, allows personal administration to students during the testing process, and saves time when scoring and analyzing the results. It also permits the direct tracking of students by displaying score reports immediately and storing them in individual logfiles in an integrated data management system (DiCerbo, Xu, Levy, Lai, & Holland, 2017). Despite the many benefits of TBA, its equivalence to traditional assessment is still a subject of debate.

Several studies conducted over the past decade have produced divergent results on the equivalence of TBA and paper-based tests (Gates & Kochan, 2015; Williamson et al., 2017). Research by Kim and Huynh (2010) demonstrated that the English-language test being measured was equivalent in terms of internal consistency across modes of administration and that most items performed similarly between the paper and online groups using differential item functioning (DIF) analysis. As for the reading test, the cross-mode equivalence was confirmed with respect to model construct reliability, and no significant difference was found for multiple-choice format as regards item difficulty (Buerger, Kroehne, Koehler, & Goldhammer, 2019). However, Schroeders and Wilhelm (2010) found that children achieved higher scores in paper versions than their peers did in computer test versions. They also found that a significant difference occurred in science subjects, but the mean scores in mathematics and social studies were not significantly different between the two modes of administration. Furthermore, a study by Hassler Hallstedt and Ghaderi (2018) concluded that although students achieved lower scores on the tablet-based version of a basic math test than they did on the paper-based one, there were consistent results in terms of validity and reliability across the two test formats. Meanwhile, findings by Neumann and Neumann (2019) also suggested that the construct validity of the tablet-based test version was consistent with that of the paper-based one, but the mean score of the students who took the tablet-based test was higher than that of the group completing the traditional version.

For cognitive tests, the transition from traditional delivery to an online platform may increase reliability and standardization (Csapó et al., 2012). Specifically, the reliability of the IR test showed a good level in both the paper and online formats but favored the online mode (Csapó, Molnár, & Nagy, 2014). Moreover, an investigation by Schroeders and Wilhelm (2010) compared the effects of differing media delivery of reasoning tests (verbal, numerical and figural). The authors found no significant differences in test reliability across these different media, but the average score on the paper test was higher than that of the other test. However, a study by Bailey, Neigel, Dhanani, and Sims (2018) did not confirm measurement invariance through the structural equation modeling approach across the computer-based and paper-based versions, although the reliability of the spatial test favored the latter. Likewise, Williamson et al. (2017) found that students tend to perform better on the spatial test in online format (effect sizes:  $d=0.27$  for the Mental Rotation Test), but they performed significantly better on both subtests on the reasoning test in the paper-based version (effect sizes:  $d=0.2$  and  $0.30$ ).

In short, the average score seems to favor the paper-based group, although the psychometric properties of the test, i.e. reliability and validity, are consistent across the two formats. Most previous empirical studies considered evaluating the validity evidence for the internal structure of a test and the proposed interpretation of the mean total scores for particular purposes (Gates & Kochan, 2015; Williamson et al., 2017). The recent study by Lemmo (2021) suggested a framework with three dimensions: content, format and solution for comparing students' performance when solving mathematics problems in paper-based and digital environments. It is essential to use multiple sources of evidence to evaluate the performance of a test. This study tends to employ a statistically multifaceted approach focusing on invariance measurement, the

performance distribution of the Rasch model scale and DIF analysis to compare equivalence between the two modes of administration.

#### 4. Research questions

This study aimed to evaluate the extent to which modes of administration affect the cognitive tests. We are interested in the differences in validity of the internal structure and scores when comparing the dual modes of administration on the same test, taking into consideration the item, task and test levels. The following three research questions guided our study:

1. Are the adapted tests suitable for students in the Vietnamese context using the two modes of administration?
2. Is there any evidence of equivalence between the online and paper-based groups on the cognitive tests at the item and task levels?
3. How do the different modes of administration impact students' performance at the test level?

#### 5. General data analysis strategy

We conducted two investigations with two kinds of cognitive tests in this study. Each test was composed in two paper-based and computer-based formats, but the contents, orders and number of items were the same in both versions. The online testing was administrated via the Electronic Diagnostic Assessment System (eDia) (Csapó & Molnár, 2019). The eDia system is a browser-based assessment platform and works sufficiently with Standard Internet browsers, such as Mozilla Firefox and Google Chrome (Greiff et al., 2018).

We involved three main analytical approaches at the item, task and test levels to evaluate the performance comparability within the two modes of administration. First, the raw score for each item was scored 1 and 0 for correct and incorrect answers, respectively. The raw data were entered by the researchers and assistant teachers for the paper-based tests, while the results of online tests were recorded automatically during the testing process in the eDia platform. We then scaled the test results with the Rasch model based on item response theory (IRT) to examine the relationship between the test-takers' abilities and their responses to the test items. The model presents the probability of a person completing an item predicted from the relationship between the person's ability level in describing a latent trait and the item's difficulty level on the same continuum (Rasch, 1960). A "logit" scale was applied to express item difficulty on a linear scale that can essentially range from negative infinity to positive infinity. We manipulated the data for dichotomous items with maximum likelihood estimation (MLE) parameters. The relationships between item difficulties and students' performance were expressed visually on a Wright map on the same linear scale. The map supports a judgement of the predicted order of item difficulty with the empirical order of item difficulty in a data set. In this study, we employed Wright maps to illustrate the distributions of students' performance in the two modes of administration.

The reliability of a measurement can be defined as whether it is consistent over time, according to Cronbach (1990). To assess measurement reliability, there are many methods available. Internal consistency measured through Cronbach's alpha or omega is one of the most popular method to determine interitem reliability (Gliner, Morgan, & Leech, 2016). These indicators are related to the validity of the construct being measured.

With regard to comparing model fit in the Rasch model, we referred three main indices, including the final deviance, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) (Wilson, De Boeck, & Carstensen, 2008). The final deviance is calculated by  $-2 \log(L)$ , with  $L$  being the maximum of the likelihood function of a given model, while the AIC and BIC are computed from the deviance:  $AIC = deviance + 2n_p$  and  $BIC = deviance + \log(N)n_p$  ( $n_p$ : the number of parameters;  $N$ : the number of persons). Based on the final deviances, AIC

and BIC, a model is estimated as a better fit model if it produces a lower result in these coefficients.

Another concept equally important when studying differences between groups is measurement invariance, which implies that measurements in different groups are comparable. It is considered as a prerequisite for comparing measurements in different groups to ensure the same measure construct across subgroups of students. Several approaches are available to measure invariance (Vandenberg & Lance, 2000). In this study, we employed DIF analysis, which allows detecting whether items should be retained on the test or revised with regard to fairness among groups. It can be used to compare students' performance at the item level in subgroups of gender or modes of administration. DIF analysis was conducted in dichotomously scored items using Angoff's delta plot method (see Angoff, 1982) with the R deltaPlotR package (Magis & Facon, 2014). The main benefits of the approach are that it is a simple and straightforward method, which relies on the particular items themselves without the necessity of intensive adding power, so it can handle even small samples of respondents without relying on asymptotic assumptions. The perpendicular distances of all delta points to the major axis are calculated. If an item is located above the major axis (large positive distances), it is estimated to be easier for respondents in the reference group, and vice versa. A small distance implies that the item difficulty is similar across groups, indicating it as a non-DIF item.

Differential bundle functioning (DBF) or item bundle DIF analysis is a natural extension of the DIF approach. The framework for a DBF often contains a bundle of items tied to the corresponding content area that matches the test specifications and the associated subtest consisting of the remaining items. In the current study, we used the framework developed by Douglas, Roussos, and Stout (1996), which used the Simultaneous Item Bias Test (SIBTEST) method (Shealy & Stout, 1993) to examine DIF in bundles. Douglas et al. (1996) proposed a multidimensionality-based DIF analysis paradigm that includes DBF. The basic principle in recognizing possible differentially functioning bundles is to identify those items through a common secondary ability in addition to the target items. We used the R mirt package (Chalmers, 2012) to test DBF with the function of the SIBTEST. The SIBTEST provides an estimate of the unidirectional DIF index ( $\beta_s$ ). A negative  $\beta_s$  indicates a DBF favoring the focal group, and a positive  $\beta_s$  implies a DBF favoring the reference group (Douglas et al., 1996; Shealy & Stout, 1993).

#### 6. Study 1: paper-based and technology-based assessment with supervision

##### 6.1. Participants

The final data set was drawn from 715 students in six public schools in a southern province in Vietnam. A total of 20 classes were recruited for this study, based on matching the equivalence of students' school

**Table 1**  
Characteristics of the current sample in the four cohorts.

Grade	N	Online/ Paper ratio (%)	Online		Paper	
			Male/ Female ratio (%)	Mean age (years)	Male/ Female ratio (%)	Mean age (years)
6	103	42.7/ 57.3	34.1/ 65.9	11.3	47.5/ 52.5	11.2
9	115	43.5/ 56.5	46.0/ 54.0	14.3	49.2/ 50.8	14.2
10	246	56.5/ 43.5	38.8/ 61.2	15.1	52.3/ 47.7	15.2
11	251	41.4/ 58.6	46.2/ 53.8	16.2	51.0/ 49.0	16.2
All	715	47.1/ 52.9	41.5/ 58.5	14.8	50.5/ 49.5	14.8

performance in the previous semester from the sample cluster list of 20 public schools. Table 1 presents the characteristics of the four student cohorts in the 6th, 9th, 10th and 11th grades.

We referred to students' school grade achievement in the previous semester to check prior ability equivalence between groups because previous studies (e.g., Díaz-Morales & Escribano, 2013; Van Vo & Csapó, 2020) had confirmed a strong relationship between IR and students' school grade performance. School grade achievement is computed using 11–13 school subjects, obtained from both elective and compulsory subjects. At the secondary education level in Vietnam, this index is scaled into five categories: excellent, good, fair, weak and poor. Students receive school performance reports at the end of every semester. Students reported their school performance in the first part of the instrument test. A Pearson's chi-square test for independence suggested that the students' school grade achievement in the previous semester shows the same distribution between the online and the paper-and-pencil groups. (Table 2).

We conducted this study in the first semester of the 2019–2020 academic year. Students were asked to participate voluntarily in the study, and results of the test served for the research purpose, not for earning credits for their school performance. For the paper-based group, a test booklet with two single items on each page was handed to each student along with an answer sheet. A teacher introduced the research aims and guided the students through the appropriate practice steps following our instructions. For the online test, each student received a link and an individual code to access the test. Two teachers were present in the computer room or classroom to observe and resolve technology issues during the testing process. Students had 30 min to complete this test and were supervised by their teachers as part of the regular school timetable in their everyday classroom (for the paper-based group) or in the computer rooms (for the online group). Students were encouraged to take notes when doing the test in both conditions.

## 6.2. Instruments

The IR test was adapted from the item pool developed by the Research Group on the Development of Competencies. These items were composed of nonverbal material to measure the IR skills in a general subject. The items were developed in Hungarian and contained four kinds of tasks involving figure series completion (FS), figure analogies (FA), number series (NS) and number analogies (NA). This test has been used in several empirical studies in cross-cultural contexts to establish its reliability and predictive validity for use with school-age populations (Csapó, Hotulainen et al., 2019; Kambeyo & Wu, 2018; Van Vo & Csapó,

2020). The basic criteria for selecting items were based on the structure of each item and the empirical evidence from previous studies. We considered the diversity of items and avoided questions that contained similar rules on the test. Item difficulty from the previous empirical studies (e.g., Csapó, Hotulainen et al., 2019; Kambeyo & Wu, 2018; Van Vo & Csapó, 2020) was also referred when selecting the items for the IR test. The test was expected to measure the appropriate abilities of the entire student sample. Ultimately, 20 items (five items for each subtest) were used in this study. To ensure the same order of presentation in both versions, we designed a fixed-length test in this study. Fig. 1 shows examples of test items available in both the paper and online versions.

## 6.3. Findings of study 1

### 6.3.1. Reliability and validity

Table 3 presents Cronbach's alpha ( $\alpha$ ), McDonald's omega ( $\omega$ ) and the Pearson correlation between the subtests. Generally, the internal consistency reliability was acceptable though not excellent for both test versions, but the paper-and-pencil format seemed somewhat better than the online one. There were significant positive correlations between the subtests, with those in the paper-based format being stronger than those on the online test, while the strongest one was found between the figure series completion and figure analogies tasks in both versions.

Table 4 summarizes the psychometric properties of the IR test comparing the online and paper groups. In classical statistics, the percentage of correct answers and discrimination values are acceptable when they are higher than 0.3 (Ebel & Frisbie, 1991) and comparable between the two groups for most test items, except item 20. The results of the Rasch model analysis showed that the test items fitted the model to the data quite well in both versions. The infit for single items (weighted mean squares, MNSQ) ranged from 0.85 to 1.28 (Mean=0.99, SD=0.09) in the online group and from 0.87 to 1.20 (Mean=1.01, SD=0.11) in the paper-based group. The item difficulty ranged from −1.79 to 2.52 and from −1.57 to 2.16 on the online and the paper-based tests, respectively. Overall, these results suggest that the models were well supported with the empirical data in both test formats. Furthermore, the unidimensional model of the online sample resulted in a final deviance of 6548.59 with 21 parameters (AIC=6590.59, BIC=6601.67), while that of the paper sample was estimated at a final deviance of 6847.79 and 21 parameters (AIC=6889.79, BIC=6901.92). Consequently, the deviance, AIC and BIC of the online group were lower than those of the paper group, suggesting that the unidimensional model of the online sample fitted better to the empirical data than that of the paper group.

**Table 2**

Distribution of the online and paper groups by school achievement in the previous semester.

Grade	Mode		Poor	Weak	Fair	Good	Excellent	Total	$\chi^2$	p
6	Online	N	0	0	1	19	24	44	1.387	0.500
		%	0.00	0.00	2.27	43.18	54.55	100		
	Paper	N	0	0	0	25	34	59		
		%	0.00	0.00	0.00	42.37	57.63	100		
9	Online	N	0	0	7	22	21	50	0.004	0.998
		%	0.00	0.00	14.00	44.00	42.00	100		
	Paper	N	0	0	9	29	27	65		
		%	0.00	0.00	13.85	44.62	41.54	100		
10	Online	N	0	0	2	65	72	139	0.240	0.887
		%	0.00	0.00	1.44	46.76	51.80	100		
	Paper	N	0	0	1	48	58	107		
		%	0.00	0.00	0.93	44.86	54.21	100		
11	Online	N	0	2	12	52	38	104	3.933	0.269
		%	0.00	1.92	11.54	50.00	36.54	100		
	Paper	N	0	0	12	74	61	147		
		%	0.00	0.00	8.16	50.34	41.50	100		
All	Online	N	0	2	22	158	155	337	2.492	0.477
		%	0.00	0.59	6.53	46.88	45.99	100		
	Paper	N	0	0	22	176	180	378		
		%	0.00	0.00	5.82	46.56	47.62	100		



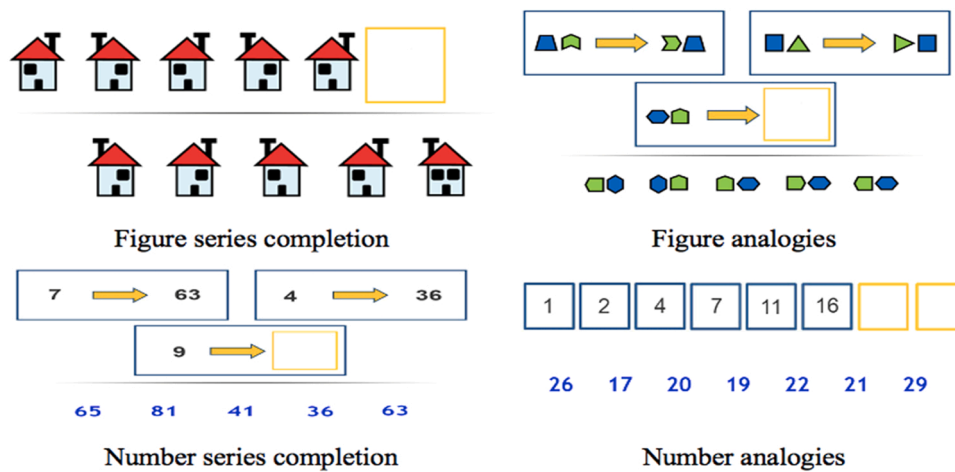


Fig. 1. Sample items on the inductive reasoning test.

Table 3

Internal consistency indicated by Cronbach's alpha ( $\alpha$ ), McDonald's omega ( $\omega$ ) and the intercorrelations for the subscales.

	Online (N = 337)					Paper (N = 378)				
	$\alpha$	$\omega$	FS	FA	NA	$\alpha$	$\omega$	FS	FA	NA
FS	0.57	0.62				0.67	0.75			
FA	0.60	0.76	0.310			0.57	0.62	0.503		
NA	0.61	0.69	0.307	0.293		0.50	0.55	0.404	0.370	
NS	0.50	0.53	0.264	0.322	0.395	0.69	0.72	0.435	0.456	0.446
All	0.76	0.80				0.83	0.85			

Note:  $p < .001$  for all correlation coefficients.

Table 4

The psychometric parameters of the IR test by mode of administration.

No.	Item	Correct answer		Discrimination		Difficulty		Infit	
		Online	Paper	Online	Paper	Online	Paper	Online	Paper
1	FS01	83.68	83.86	0.39	0.62	-0.84	-0.56	1.01	0.85
2	FS02	87.83	86.77	0.32	0.58	-1.23	-0.85	1.04	0.86
3	FS03	82.49	83.33	0.43	0.55	-0.74	-0.52	0.98	0.88
4	FS04	92.28	92.33	0.37	0.42	-1.79	-1.56	0.89	0.98
5	FS09	24.63	37.83	0.40	0.33	2.43	2.16	1.01	1.14
6	FA02	91.10	91.53	0.37	0.37	-1.62	-1.44	0.94	1.06
7	FA05	59.94	61.11	0.52	0.53	0.62	0.95	0.96	0.96
8	FA06	59.94	54.76	0.52	0.47	0.62	1.28	0.96	1.05
9	FA07	74.78	82.28	0.48	0.52	-0.20	-0.42	0.95	0.95
10	FA10	64.09	76.19	0.35	0.50	0.40	0.04	1.12	0.99
11	NA01	75.37	82.54	0.50	0.35	-0.24	-0.45	0.96	1.15
12	NA02	82.49	89.95	0.53	0.51	-0.75	-1.22	0.87	0.91
13	NA03	85.76	86.77	0.38	0.41	-1.03	-0.85	0.96	1.06
14	NA05	80.12	80.42	0.55	0.50	-0.56	-0.27	0.87	1.04
15	NA07	33.53	38.10	0.28	0.25	1.92	2.15	1.20	1.28
16	NS01	87.54	90.21	0.50	0.59	-1.20	-1.25	0.93	0.93
17	NS02	72.40	68.31	0.50	0.56	-0.06	0.48	0.96	1.00
18	NS03	60.53	70.11	0.45	0.50	0.59	0.44	1.02	1.07
19	NS05	48.66	59.52	0.42	0.56	1.17	1.03	1.05	0.98
20	NS07	23.15	62.96	0.22	0.53	2.52	0.85	1.18	0.98

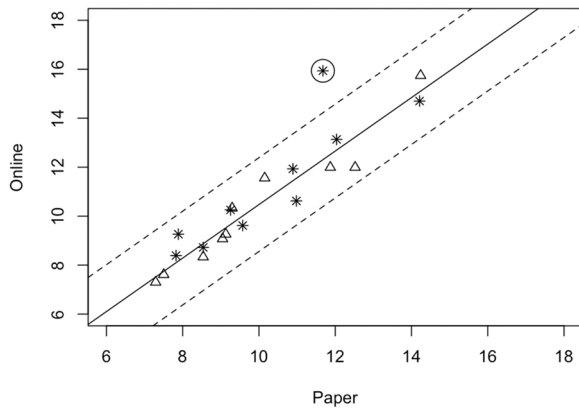
### 6.3.2. A comparison by DIF analysis

DIF analysis was conducted with the online group set as a focal group. DIF analysis using Angoff's delta method with item purification parameters of the major axis ( $a = -0.4523$ ,  $b = 1.0925$ ) in the last iterations and a detection threshold of 1.29 (significance level: 5 %) suggested that 19 items displayed no DIF and only item 20 (NS07) was detected as a DIF item, these results thus favoring the paper group. Interestingly, most (7 out of 10) of the items with figure-related material yielded a positive effect size, while most (6 out of 10) of the items with number-related material had a negative effect size. In other words, the

items comprising figures (represented by triangles) seemed to favor the online group, while the items containing numbers (represented by stars) tended to favor the paper group, as depicted in Fig. 2.

### 6.3.3. DBF analysis for comparison between the two modes of administration

For DBF analysis, the online group was also used as the focal group. Table 5 summarizes the results of the SIBTEST method at the task (item bundle) level. DBF analysis determined that significant DBF for the bundle of items was found in the figure analogies and figure series



**Fig. 2.** Delta plots for the dual modes of administration of the IR test. Note: The items are located above the major axis, indicating that they are easier for the respondents in the reference group (paper group).

**Table 5**  
Summary of the SIBTEST results by task level.

Item bundle	No. of items	$\beta_s$	p-value	Result favors
Figure analogies	5	-0.317	0.000	Online
Figure series completion	5	-0.455	0.000	Online
Number analogies	5	0.029	0.762	No DBF
Number series completion	5	0.095	0.078	No DBF

completion tasks, with the online group being favored. However, no DBF was indicated in the item bundle of the number series completion and number analogies tasks, although it seemed to favor the paper group.

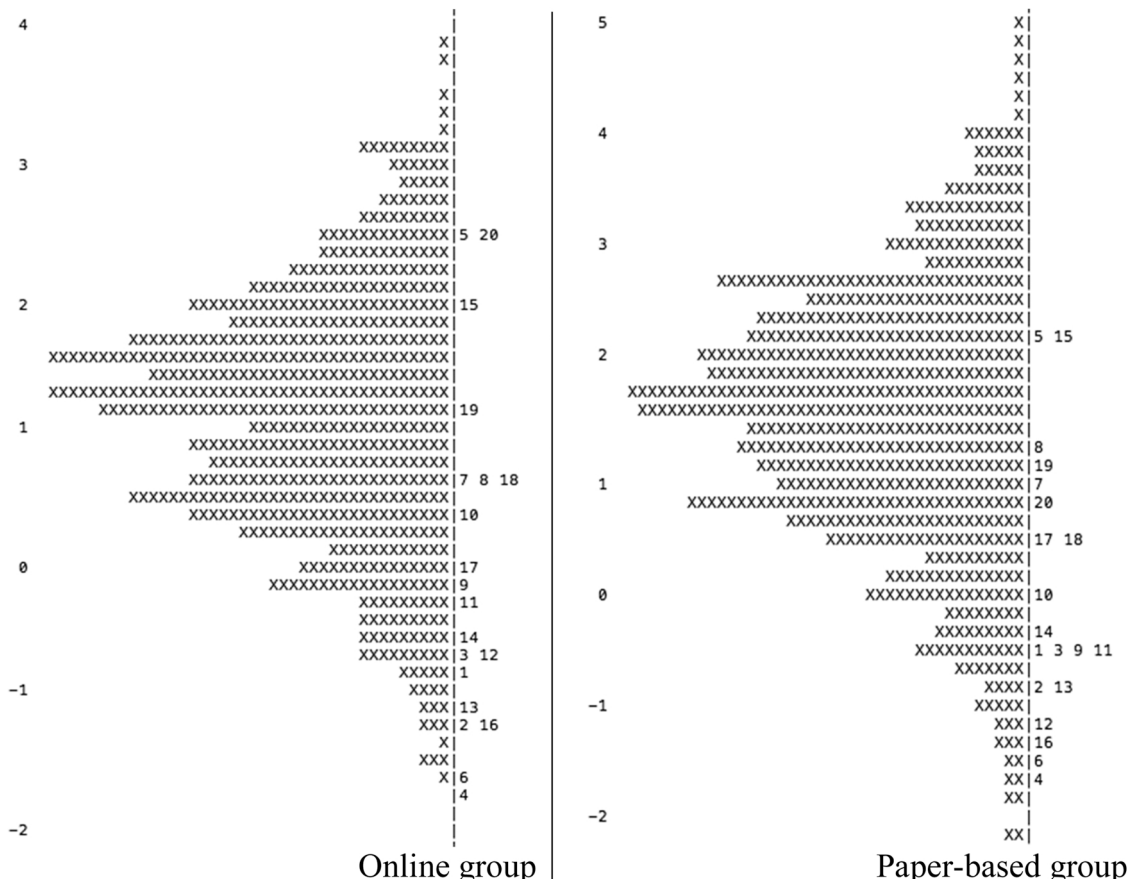
Generally, DBF analysis showed that the performance of the online group on the bundle of figure-related items was better than that of the paper-based group, whereas the results were apparently reversed in the number-related items.

#### 6.3.4. A descriptive comparison of the students' performance

The average proficiency of the participants in the online group was 1.14 (SD=1.19) on the MLE scale of the Rasch measurement model, and that of the paper group was 1.56 (SD=1.39). In comparison with average item difficulty, the students' proficiency was estimated as higher than item difficulty (set at 0 logit). The students in the paper-based group found the test a bit easier than those in the online group. On average, the online participants completed 13.7 out of 20 items correctly (68.5 %), while the students in the paper-based group did 14.8 out of 20 items correctly (74.4 %).

The Wright maps in Fig. 3 present the patterns of the students' performance in the two groups. The results from the Wright map also suggest that the participants using the paper-based test medium tended to complete item 20 easily, as the results indicated the location of this item in the middle of the scale. However, this was not the case with the online participants because the map suggests that that item tended to be the most difficult one for them as it stood at the top of the scale. The finding is consistent with the result of Delta plots analysis, as discussed above. All items covered most of the participants' skills, but, in general, the test seemed easy for the students in both groups. This might be explained by the fact that almost all students had Fair, Good or Excellent school performance. All in all, the students' achievement on the online version was a bit lower than that of their peers on the paper-based version.

Furthermore, we conducted a *t*-test to examine the difference in performance between the two groups. Table 6 provides a brief account



**Fig. 3.** The Wright maps for the online and paper-based groups. Note: Each "x" represents 0.6 cases.

**Table 6**

Students' performance on the IR test by mode of administration.

Grade	Online		Paper		<i>t</i>	<i>p</i>	Cohen's <i>d</i>
	N	Mean (SD)	N	Mean (SD)			
6	44	1.07(1.04)	59	−0.03 (1.01)	5.39	<0.001	1.07
9	50	0.88(1.23)	65	1.53(1.23)	−2.78	0.006	0.52
10	139	1.04(1.18)	107	1.68 (1.27)	−4.02	<0.001	0.52
11	104	1.41(1.21)	147	2.12(1.18)	−4.55	<0.001	0.59
All	337	1.14 (1.19)	378	1.56 (1.39)	−4.33	<0.001	0.32

of the students' performance on the test in the two modes of administration grouped by grade level. Generally, the average score of the paper-based group was higher than the mean score of the online group using Cohen's effect size value ( $d=0.32$ ), suggesting a small to medium significance. All the cohorts in the paper-based group scored higher than those in the online group, except in the 6th grade, where the students did the online test better than their peers who took the paper-based test.

## 7. Study 2: technology-based and paper-based testing without supervision

### 7.1. Participants

The final data of this study was collected from 731 students from the 8th to 12th grades in nine secondary schools in Vietnam. As shown in Table 7, the mean ages of the students in the online and paper groups were 15.3 years and 15.5, respectively, and the male-to-female ratio seemed equivalent in each cohort and the whole sample. The students participated in the study voluntarily after the teachers introduced the aims of our research project. They were encouraged to perform the test independently in no more than 30 min, which served for the research purpose without impacting their school scores in the case. Because of the COVID-19 pandemic, most traditional schooling activities were restricted, and thus the students took the test individually at home either in paper-based or online format without teacher supervision. Most of the data were collected during March and April 2020.

For the purposes of this study, we referred to physics grades in the previous semester to examine the equivalence of the prior abilities among the students in the two groups. Previous studies (Hejnová, Eisenmann, Cihlár, & Příbyl, 2018; Schwichow, Osterhaus, & Edelsbrunner, 2020) have indicated that CVS and content knowledge in physics are closely tied, so we considered the results of the final test in physics as an index to support the assumption of prior ability equivalence between the two groups. Table 8 presents the summarized results of the *t*-test for comparing physics grades in the previous semester between the students in the online group and those in the paper group. Overall, there was no significant difference between the two groups in

**Table 7**

Characteristics of the participants in study 2.

Grade	N	Online/ Paper ratio (%)	Online		Paper	
			Male/ Female ratio (%)	Mean Age (years)	Male/ Female ratio (%)	Mean Age (years)
8	150	52.0/ 48.0	54.7/ 45.3	13.1	50.0/ 50.0	13.6
9	144	50.7/ 49.3	50.7/ 49.3	14.6	50.7/ 49.3	14.6
10	235	48.9/ 51.1	47.3/ 52.7	15.8	50.0/ 50.0	15.8
11	129	31.0/ 69.0	33.3/ 66.7	16.7	29.3/ 70.7	16.8
12	73	58.9/ 41.1	55.2/ 44.8	17.7	61.4/ 38.6	17.8
All	731	47.7/ 52.3	42.1/ 57.9	15.3	41.9/ 58.1	15.5

**Table 8**

Comparison of students' achievement in physics in the previous semester.

Grade	Online		Paper		<i>t</i>	<i>p</i>
	N	Mean (SD)	N	Mean (SD)		
8	78	8.38(1.29)	72	8.13(1.35)	1.16	0.248
9	73	8.18(1.20)	71	7.86(1.18)	1.64	0.102
10	115	7.85(1.57)	120	7.60(1.69)	−0.17	0.247
11	40	7.93(1.53)	89	7.91(1.28)	0.07	0.947
12	43	7.73(1.72)	30	8.32(1.14)	−1.74	0.086
All	349	8.03(1.47)	382	7.88(1.42)	1.45	0.146

terms of physics achievement in the separate cohorts or in the entire sample.

### 7.2. Instruments

The control of variables in physics (CVSP) test contained 24 items (eight items for each sub-skill task) to assess CVS in three sub-skills (*identifying* controlled experiments, *interpreting* the outcome of a controlled experiment and *understanding* the determinacy of confounded experiments). The items were adapted from Schwichow et al. (2016), the American Association for the Advancement of Science AAAS, (2012) and TIMSS (1997), with other new items developed by the authors. The content of items relates to the basic physics concepts (mechanics, heat and thermodynamics, and electricity and electromagnetism) in the secondary educational program in Vietnam. The items were formatted in the multiple-choice style, consisting of a stem with one correct answer and three distractors. We visualized graphic representations to minimize the influence of students' varying reading ability levels. Test validity was confirmed in the previous study (Van Vo & Csapó, 2021b). The test was designed in fixed-length style, in which the items and their orders were presented in the same way in both versions. Fig. 4 illustrates a sample item on the CVSP test.

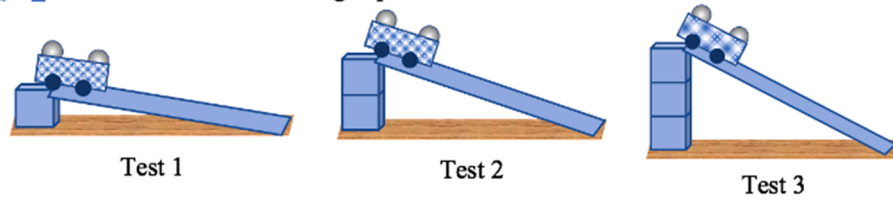
### 7.3. Findings of study 2

#### 7.3.1. Reliability and validity

Cronbach's alpha ( $\alpha$ ), McDonald's omega ( $\omega$ ) and the Pearson correlation between sub-skill tasks are shown in Table 9. Both tests demonstrated an adequate level in terms of internal consistency reliability, although the students performed somewhat better in online testing when compared to the paper-based version. A significant positive correlation was found between the subtests. The strongest correlation was found between the *identifying* and *interpreting* tasks, followed by that between the *interpreting* and *understanding* tasks on both versions of the test.

Table 10 summarizes the psychometric properties of the CVSP test in the online and paper-based formats. As suggested by Ebel and Frisbie (1991), the discrimination indices were comparable for both test versions, but six items in the online version and four items in the paper one were still less than 0.3. Rasch model analysis also suggested that the model fitted the data well at item level for both test versions. The infit for single items ranged from 0.85 to 1.24 (Mean=1.00, SD=0.12) and from 0.79 to 1.35 (Mean=1.00, SD=0.12) for the online and the

Q16\_IN04. Tom did the following experiment:



What can he conclude from these experiments?

- A. The mass of the cart affects how the cart performs.
- B. The height of the ramps affects how the cart performs.
- C. The cart's mass affects how the cart performs, and the height of the ramp affects how the cart performs.
- D. It is not possible to reach any valid conclusion.

Fig. 4. A sample item of the *Interpreting* sub-skills task type on the CVSP test.

Table 9

Internal consistency indicated by Cronbach's alpha ( $\alpha$ ), McDonald's omega ( $\omega$ ) and the intercorrelations for the subscales.

Sub-skill	Online (N = 337)				Paper (N = 378)			
	$\alpha$	$\omega$	Identifying	Interpreting	$\alpha$	$\omega$	Identifying	Interpreting
Identifying	0.64	0.73			0.67	0.71		
Interpreting	0.69	0.75	0.608***		0.61	0.68	0.602***	
Understanding	0.52	0.63	0.472***	0.555***	0.48	0.58	0.477***	0.516***
All	0.81	0.84			0.80	0.82		

Note: \*\*\* $p < .001$

Table 10

The psychometric characteristics of the CVSP test by mode of administration.

No.	Item	Correct answer		Discrimination		Difficulty		Infit	
		Online	Paper	Online	Paper	Online	Paper	Online	Paper
1	ID01	63.39	56.01	0.35	0.50	-0.97	-0.88	1.03	0.94
2	ID03	40.71	27.44	0.27	0.30	0.10	0.61	1.13	1.08
3	IN02	31.42	25.17	0.23	0.19	0.56	0.71	1.20	1.22
4	ID02	55.74	56.24	0.52	0.54	-0.61	-0.99	0.96	0.98
5	ID04	72.40	65.76	0.35	0.43	-1.48	-1.46	1.08	1.00
6	UN08	39.07	43.08	0.19	0.36	0.21	-0.32	1.22	1.12
7	IN05	39.62	25.62	0.36	0.42	0.20	0.71	1.08	0.99
8	IN07	39.07	37.64	0.58	0.44	0.20	-0.03	0.87	1.04
9	ID05	72.40	59.64	0.53	0.52	-1.50	-1.16	0.90	0.91
10	UN02	57.92	44.67	0.53	0.54	-0.71	-0.39	0.94	0.91
11	ID06	65.30	49.89	0.55	0.66	-1.06	-0.60	0.89	0.79
12	IN04	36.34	30.61	0.55	0.54	0.37	0.32	0.94	0.89
13	UN03	7.10	7.94	0.14	0.20	2.60	2.26	1.03	0.99
14	ID07	45.36	46.94	0.42	0.40	-0.12	-0.53	1.03	1.05
15	IN01	40.16	44.22	0.47	0.44	0.15	-0.44	1.01	1.05
16	IN03	55.46	43.99	0.60	0.59	-0.62	-0.29	0.90	0.86
17	UN06	21.58	19.50	0.52	0.58	1.20	1.03	0.85	0.83
18	IN08	68.31	54.20	0.52	0.54	-1.27	-0.82	0.92	0.91
19	UN07	3.55	7.48	0.27	0.14	3.47	2.34	0.98	0.96
20	UN01	51.37	43.31	0.58	0.32	-0.38	-0.21	0.89	1.13
21	UN04	29.78	20.18	0.39	0.39	0.67	1.03	1.02	0.98
22	ID08	52.46	41.04	0.52	0.51	-0.48	-0.29	0.92	0.94
23	UN05	66.94	64.40	0.11	0.02	-1.16	-1.28	1.24	1.35
24	IN06	30.33	24.72	0.57	0.39	0.62	0.68	0.85	1.03

paper-based version, respectively. The item difficulty ranged from -1.50 to 3.47 in the online format and from -1.46 to 2.34 in the paper-based format. Moreover, the final deviance in the unidimensional model of the online sample is 9479.76 with 25 parameters estimated ( $AIC=9529.76$ ,  $BIC=9543.33$ ), and this value in the paper sample is 10,372.01 with 25 parameters estimated ( $AIC=10,422.01$ ,  $BIC=10,436.56$ ). It is seen clearly that the deviance, AIC and BIC in the paper group were greater than those in the online group, suggesting that the unidimensional model fitted better to the online data than the paper

one.

### 7.3.2. Measurement invariance with DIF analysis

Fig. 5 depicts the results of DIF analysis applying Angoff's delta method when the online group was defaulted as a focal group. Examination of DIF operated with item purification parameters of the major axis:  $a = -3.04$ ,  $b = 1.17$ , with a detection threshold of 1.03 and a significance level of 5 %. The findings showed that there was no DIF item that was detected as a DIF item. Nevertheless, ten items had a negative



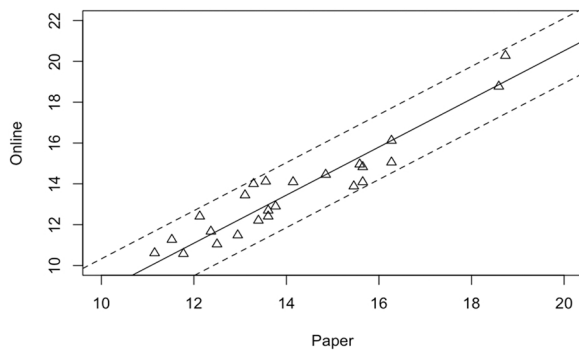


Fig. 5. Delta plots for the dual modes of administration of the CVS test.

effect size, implying that the results favor the paper group, while fourteen items were estimated with a negative effect size, suggesting that they were a bit easier for the online group.

### 7.3.3. A comparison by DBF analysis

The results of the DBF analysis were plotted in Table 11 when the online group was considered as the focal group with the SIBTEST method. Surprisingly, a significant DBF was indicated in all the bundles of CVS sub-skill items, with the results favoring the online group. In other words, the students who took the online test appeared to perform better than their peers who used the paper-based format in all the item bundles of the *identifying*, *interpreting* and *understanding* sub-skills.

### 7.3.4. Comparison of students' performance in dual media delivery modes

The online participants had an average proficiency of  $-0.33$  digits ( $SD = 1.13$ ) and that of the paper-based group was  $-0.54$  ( $SD = 1.16$ ). In comparison to item difficulty (Mean =  $0.00$ ,  $SD = 1.00$ ), students' proficiency was estimated lower, suggesting that students felt the SVP test was somewhat difficult and that the paper group found the test more difficult than the online group did. On average, the online participants completed 10.89 out of 24 items (45.38 %) correctly, while the paper group managed 9.63 items (40.13 %).

Additionally, the Wright maps in Fig. 6 illustrate the contribution of students' performance in the two groups in relation to item difficulty. Though a general overview did not show significant differences between the two patterns on the maps, more online participants scored higher than 0 (digits). Items 13 and 19 were supposed to be the most difficult ones in both versions. Specifically, item 19 fell out of the spectrum of online test-takers. The locations of the other items covered the respondents' proficiency well, but the test seemed somewhat difficult for both student groups in general. Overall, students' achievement in the online version was higher than that for students who took the paper-based test.

Furthermore, the *t*-test was performed to compare students' performance between the two groups. Table 12 summarizes the results of the *t*-test on the CVSP test with regard to the two delivery modalities across the grade levels. Overall, the online group scored higher than the paper group with a small effect size with Cohen's value of 0.20. All the younger cohorts (the 8th, 9th and 10th grades) obtained significantly higher scores on the online version than their peers who took the traditional paper-based test. However, no significant disparity was found between the two 11th-grade groups, and even the paper group performed significantly higher than the online group in the 12th grade.

Table 11  
Summary of the SIBTEST results by sub-skill.

Item bundle	No. of items	$\beta_s$	p-value	Result favors
Identifying	8	$-0.709$	0.000	Online
Interpreting	8	$-0.654$	0.000	Online
Understanding	8	$-0.248$	0.006	Online

### 7.3.5. Additional analyses

The current version of the eDia platform allows us to measure the response time for each task as well as for a whole test. The average time to complete the test was 26.4 mins ( $SD=10.5$  mins). Fig. 7 depicts the mean response time (seconds) in which online participants completed each item. The response time for most of the items (19 out of 24) ranged from 30 to 60 s, and two items required more than 60 s, while there were three items which students just took under 30 s each to handle. It seems that the average response time for an item in the *understanding* task was lower than other ones (see box plot in Fig. 7). However, the result of the analysis of variance showed that response time did not significantly affect the execution of sub-skill tasks ( $F(2, 21)= 2.98$ ,  $p = .073$ ).

## 8. Discussion and conclusions

When tests evolve from paper-based assessment to TBA, equivalence is often expected no matter what the modes of administration are. Our study attempted to implement many-sided analyses to compare the performance of students across different grade levels, taking both traditional paper-and-pencil and online formats of the IR test at the item, task and test levels into consideration. As regards the average total score, the results seemed to be more supportive of the paper-and-pencil version under teachers' supervision because the students performed better than their peers did via online assessment on the same test. These results were corroborated by the findings of previous studies (e.g., Schroeders & Wilhelm, 2010; Williamson et al., 2017), which concluded that the average performance of students on the IR test was lower with digital forms of testing (e.g., using notebooks, tablets and smartphones) than with a traditional paper-based format. On the other hand, without supervision, the results seemed to favor the online format as the students who did the online testing performed better than their friends who participated in the paper-based version.

Interestingly, across delivery conditions, the results demonstrated that either with or without supervision, measurement with the Rasch model indicated that the online assessment had a better fit to the empirical data than the paper one. The results of the measurement invariance in DIF analysis with Angoff's delta approach showed the internal validity of the tests, thus demonstrating that they are acceptably comparable across the two modes of administration. The reliability and validity of the two tests were equivalent regardless of media delivery modalities. These findings are in line with those of previous studies (e.g., Csapó et al., 2014; Hassler Hallstedt & Ghaderi, 2018; Schroeders & Wilhelm, 2010). However, item 20 (NS07) was identified as a DIF item with regard to the modes of administration in the first study. Possible reasons include the location of the item or students' fatigue or boredom. Item 20 is the last item ordered in the IR test and one of the most difficult items in the number series completion subtest, which might in turn make students lose interest in attempting the last one within a series of numbers on the screen. As a result, they might guess the answer rather than completing it with their own abilities and efforts. In addition, difficulty-based item order did not impact the item statistics on a paper-and-pencil multiple-choice test (e.g., difficulty, discrimination and point biserial) (Sad, 2020), but the item order may influence item properties on an online test. This issue raises an exciting issue for further research.

Furthermore, by employing the dimensionality-based approach to identifying clusters of items, we found that DBF favored the online group in the bundle of figure-related items. DBF analysis provided an insight into the tasks when constructing the tests. In the case of these cognitive tests, the study showed that students appeared to feel it was easier to complete tasks using pictures or graphs than those with simply numerical elements and words on the online test. This provided clear evidence of a link between test modality and the materials that made up the test items. It further illuminates how TBA can be more beneficial when items are composed of visually rich materials.

As regards the age groups of the respondents, the technology-based

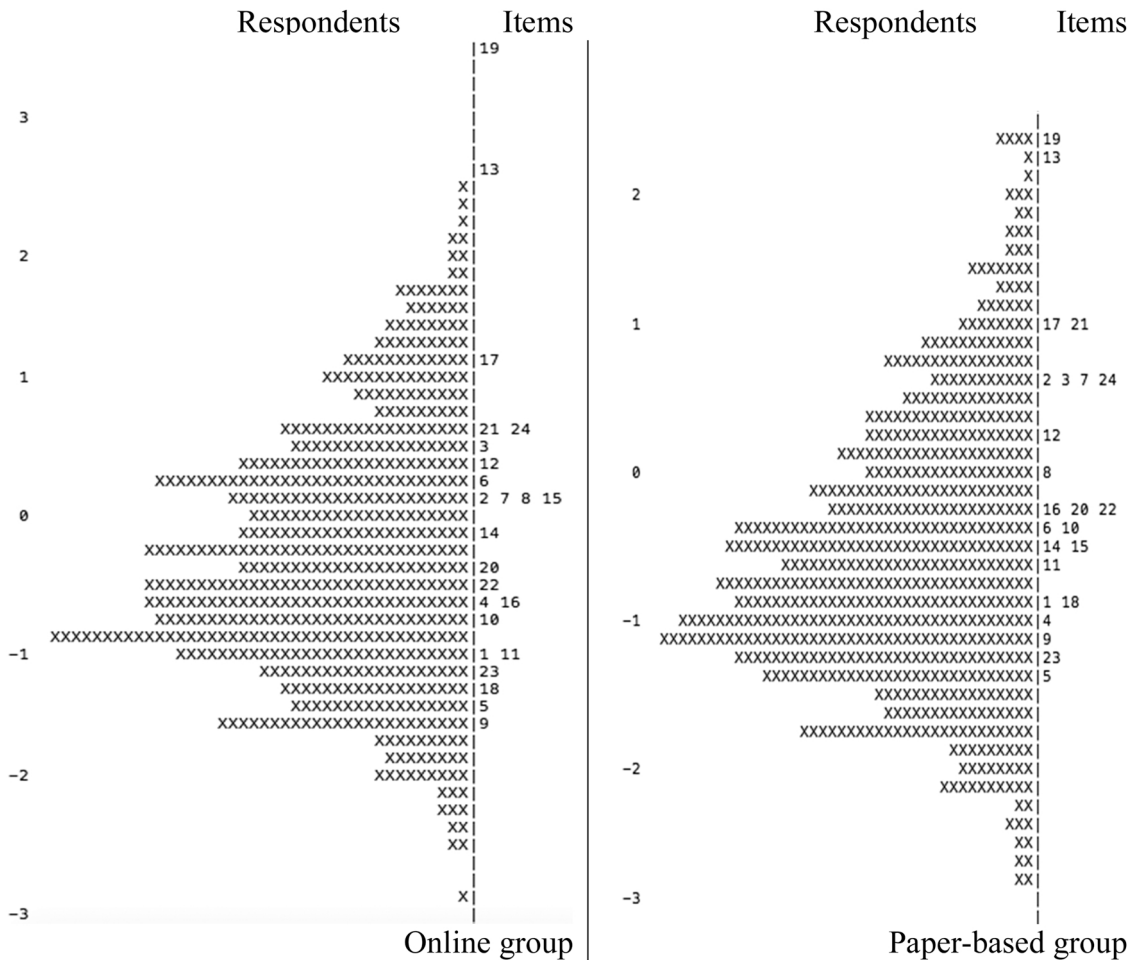


Fig. 6. The Wright maps for the online group and the paper-based group on the CVS test Note: Each “x” represents 0.6 cases.

Table 12

Comparison of the students’ performance on the CVSP test by delivery modality.

Grade	Online		Paper		<i>t</i>	<i>p</i>	Cohen’s <i>d</i>
	N	Mean (SD)	N	Mean (SD)			
8	78	−0.71(0.83)	72	−1.34(0.71)	4.96	< 0.001	0.81
9	73	−0.69(1.13)	71	−1.14(0.61)	3.01	0.003	0.50
10	115	−0.15(1.07)	120	−0.74(0.82)	4.71	< 0.001	0.62
11	40	0.19(1.25)	89	0.26(1.02)	0.07	0.758	0.03
12	43	0.21(1.21)	30	1.20(1.37)	−3.18	0.002	0.77
All	349	−0.30(1.13)	382	−0.54(1.16)	2.80	0.005	0.20

format seemed to favor the younger cohorts over the older groups, while students at the upper grade levels who took the paper-based test clearly achieved higher scores than their peers who participated in the online tests. This may derive from the levels of digital competence in the younger and older generations, with the former enjoying the benefit of stronger digital skills than the latter (Oblinger & Oblinger, 2005). Such competence can affect students’ performance on online format tests, since they are more familiar with the online format than their seniors.

Response time is an additional advantage of technology-based assessment. Investigation of response time enables us to clarify the solution behavior of test-takers that might effect bias in test performance (Wise & Kuhfeld, 2021). In this study, it seems that the items (UN03, UN05 and UN07) with an average response time of under 30 s have a low discrimination of less than 0.3 each. However, more studies should be conducted to examine the effects of time response on the psychometric properties of the items.

Some limitations are acknowledged in the present study. The data were drawn from two small samples in a province in Vietnam. When the first wave of the COVID-19 pandemic occurred unexpectedly, readiness for remote learning may have depended on individual conditions, in which students’ ICT familiarity may have influenced their performance on the online tests. Unfortunately, the current study did not cover this issue. The identification of the equivalence of students’ prior abilities was based on their school achievement in the previous semester which might not fully cover students’ cognitive proficiency. Specifically, the results from Pearson’s chi-square test (Study 1) and *t*-test (Study 2) supported the assumption of random equivalence between groups. However, students in the paper group in the first study were in a slightly higher school grade, while students who took the online test in the second investigation had a higher mean score on the physics test than the paper group. This could possibly have affected the mean score comparison between the two delivery modalities in the later analyses.

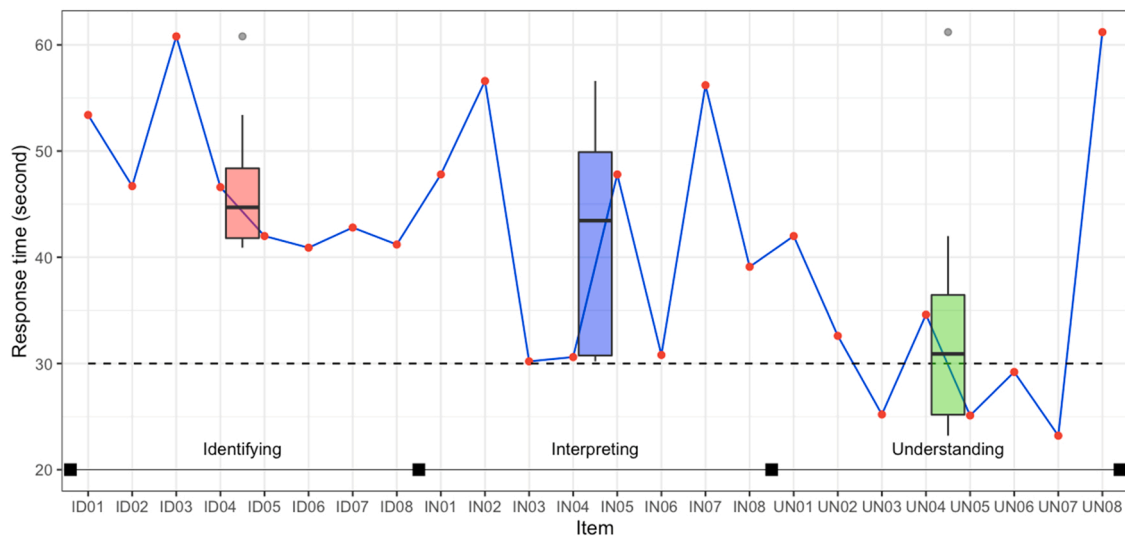


Fig. 7. The response time for each item and box plots for item bundles by sub-skill.

Although we used two different tests, these tests naturally measured students' cognitive abilities. In fact, IR test assessed students' general cognitive ability, while CVSP examined their cognitive proficiency related to specific domain of science. Therefore, in principle, the comparison could be made on the students' performance on these tests in light of the impact of the different supervision conditions, but we also acknowledged some possible bias in this case. Another concern may come from internal consistency estimates, since Cronbach's alpha and McDonald's omega were not really good, especially for individual sub-tests. Therefore, generalizations of the results should be made with caution, and further research needs to involve more studies to handle these problems.

Despite the fact that the present research is limited in terms of the comparison between online and paper-and-pencil methods as regards cognitive tests, the study employed various technical analyses that are in line with the development of educational testing. This comparability is helpful for test developers because the test can be identified as never, sometimes or always reasonably comparable at the test, task and item levels. Examining the fairness of the different modes of administration is useful for test developers and school managers when designing future tests that can provide equal opportunities regardless of the different means of delivery.

Although the investigations led to inconsistent findings regarding students' scores on the tests, the results showed the potential advantages of a technology-rich assessment when considering the psychometric characteristics of the tests, especially students' self-assessment process at home. Across the two studies, the psychometric properties of the tests and individual items were comparable in dual administration modes and even leaning toward online versions. The study also provided multifaceted approaches to examining the testing equivalence and evidence for the feasibility of transferring from paper-and-pencil to TBA in the Vietnamese context. Future research needs to consider possible factors relevant to test modality, familiarity with electronic devices (Schroeders & Wilhelm, 2010) and administration procedures. More sophisticated comparative studies should continue to explore this as advanced electronic devices are becoming essential tools both at school and home. The issue of test fairness will continue to be considered as an important construct in the upcoming decade because "we have no clear criteria on which to decide whether psychological testing is better than technology-driven profiling and prediction" (Ilescu & Greiff, 2019, p. 151). Furthermore, the COVID-19 pandemic is currently a significant challenge for traditional modes of delivery.

## Funding

This research received no specific grant from any funding agency. Open access funding provided by University of Szeged [5889].

## Conflict of interest

The authors have stated no potential conflict of interest.

## Acknowledgments

The first author of this article is a recipient of the Hungarian government's Stipendium Hungaricum Scholarship in collaboration with the Vietnamese government. The authors would like to thank the teacher's assistants for their help in coordinating and administering the instruments.

## References

- Adey, P., & Csapó, B. (2012). Developing and assessing scientific reasoning. In B. Csapó, & G. Szabó (Eds.), *Framework for diagnostic assessment of science* (pp. 17–53). Nemzeti Tankönyvkiadó.
- Adey, P., Csapó, B., Demetriou, A., Hautamäki, J., & Shayer, M. (2007). Can we be intelligent about intelligence? Why education needs the concept of plastic general ability. *Educational Research Review*, 2(2), 75–97. <https://doi.org/10.1016/j.edurev.2007.05.001>
- American Association for the Advancement of Science (AAAS). (2012). AAAS Science Assessment - Project2061. (<https://www.aaas.org/programs/project-2061>).
- Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berck (Ed.), *Handbook of methods for detecting item bias* (pp. 96–116). Baltimore, MD: Johns Hopkins University Press.
- Bailey, S. K. T., Neigel, A. R., Dhanani, L. Y., & Sims, V. K. (2018). Establishing measurement equivalence across computer- and paper-based tests of spatial cognition. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 60(3), 340–350. <https://doi.org/10.1177/0018720817747731>
- Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Studies in Educational Evaluation*, 62(April), 1–9. <https://doi.org/10.1016/j.stueduc.2019.04.005>
- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6). <https://doi.org/10.18637/jss.v048.i06>
- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, 70(5), 1098–1120. <https://doi.org/10.1111/1467-8624.00081>
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). HarperCollins.
- Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). echnological issues for computer-based assessment. *Assessment and Teaching of 21st Century Skills* (pp. 143–230). Springer Netherlands. [https://doi.org/10.1007/978-94-007-2324-5\\_4](https://doi.org/10.1007/978-94-007-2324-5_4)
- Csapó, B., Hotulainen, R., Pásztor, A., & Molnár, G. (2019). Az indukzív gondolkodás fejlődésének összehasonlító vizsgálata: online felmérések Magyarországon és

- Finnországban [A comparative study of the development of inductive thinking: online surveys in Hungary and Finland]. *Neveléstudomány [Educational Science: Education Research Innovation]*, 7(3–4), 5–24.
- Csapó, B., & Molnár, G. (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia system. *Frontiers in Psychology*, 10(JULY). <https://doi.org/10.3389/fpsyg.2019.01522>
- 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>
- Díaz-Morales, J. F., & Escribano, C. (2013). Predicting school achievement: The role of inductive reasoning, sleep length and morningness-eveningness. *Personality and Individual Differences*, 55(2), 106–111. <https://doi.org/10.1016/j.paid.2013.02.011>
- DiCerbo, K. E., Xu, Y., Levy, R., Lai, E., & Holland, L. (2017). Modeling student cognition in digital and nondigital assessment environments. *Educational Assessment*, 22(4), 275–297. <https://doi.org/10.1080/10627197.2017.1382343>
- Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, 33(4), 465–484. <https://doi.org/10.1111/j.1745-3984.1996.tb00502.x>
- Ebel, R. L., & Frisbie, D. A. (1991). Vol. 11, Issue 2. *Essentials of educational measurement*. Prentice-Hall.
- Gates, N. J., & Kochan, N. A. (2015). Computerized and on-line neuropsychological testing for late-life cognition and neurocognitive disorders. *Current Opinion in Psychiatry*, 28(2), 165–172. <https://doi.org/10.1097/YCO.0000000000000141>
- Gliner, J. A., Morgan, G. A., & Leech, N. L. (2016). *Research methods in applied settings: An integrated approach to design and analysis*. Routledge.
- Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers & Education*, 126, 248–263. <https://doi.org/10.1016/j.compedu.2018.07.013>
- Hassler Hallstedt, M., & Ghaderi, A. (2018). Tablets instead of paper-based tests for young children? Comparability between paper and tablet versions of the mathematical Heidelberg Rechen Test 1-4. *Educational Assessment*, 23(3), 195–210. <https://doi.org/10.1080/10627197.2018.1488587>
- Hejnová, E., Eisenmann, P., Cihlár, J., & Příbyl, J. (2018). Relations between scientific reasoning and culture of problem solving. *Journal on Efficiency and Responsibility in Education and Science*, 11(2), 38–44. <https://doi.org/10.7160/eriesj.2018.110203>
- Iliescu, D., & Greiff, S. (2019). The impact of technology on psychological testing in practice and policy: What will the future bring. *European Journal of Psychological Assessment*, 35(2), 151–155. <https://doi.org/10.1027/1015-5759/a000532>
- Kambeyo, L., & Wu, H. (2018). Online assessment of students' inductive reasoning skills abilities in Oshana region, Namibia. *International Journal of Educational Sciences*, 21, 1–12. <https://doi.org/10.258359/KRE-86>
- Kim, D. H., & Huynh, H. (2010). Equivalence of paper-and-pencil and online administration modes of the statewide English test for students with and without disabilities. *Educational Assessment*, 15(2), 107–121. <https://doi.org/10.1080/10627197.2010.491066>
- Kinshuk, Lin, T., & McNab, P. (2006). Cognitive trait modelling: The case of inductive reasoning ability. *Innovations in Education and Teaching International*, 43(2), 151–161. <https://doi.org/10.1080/14703290600650442>
- Lemmo, A. (2021). A tool for comparing mathematics tasks from paper-based and digital environments. *International Journal of Science and Mathematics Education*, 19(8), 1655–1675. <https://doi.org/10.1007/s10763-020-10119-0>
- Maftuh, B. (2011). Status of ICT integration in education in Southeast Asian countries. *Innovation of Classroom Teaching and Learning through Lesson Study*, 1, 1–9.
- Magis, D., & Facon, B. (2014). deltaPlotR: An R package for differential item functioning analysis with Angoff's delta plot. *Journal of Statistical Software, Code Snippets*, 59, 1.
- Mohamadi, Z. (2018). Comparative effect of online summative and formative assessment on EFL student writing ability. *Studies in Educational Evaluation*, 59(July 2017), 29–40. <https://doi.org/10.1016/j.stueduc.2018.02.003>
- Neumann, M. M., & Neumann, D. L. (2019). Validation of a touch screen tablet assessment of early literacy skills and a comparison with a traditional paper-based assessment. *International Journal of Research & Method in Education*, 42(4), 385–398. <https://doi.org/10.1080/1743727X.2018.1498078>
- Nikou, S. A., & Economides, A. A. (2018). Mobile-based assessment: A literature review of publications in major referred journals from 2009 to 2018. *Computers & Education*, 125(2018), 101–119.
- Oblinger, D. G., & Oblinger, J. L. (2005). *Educating the next generation*. EDUCAUSE. <https://www.educause.edu/ir/library/pdf/pub7101.pdf>
- Perret, P. (2015). Children's inductive reasoning: Developmental and educational perspectives. *Journal of Cognitive Education and Psychology*, 14(3), 389–408.
- Rasch, G. (1960). *Probabilistic model for some intelligence and achievement tests*. Danish Institute for Educational Research.
- Şad, S. N. (2020). Does difficulty-based item order matter in multiple-choice exams? (Empirical evidence from university students. *Studies in Educational Evaluation*, 64 (September 2019), Article 100812. <https://doi.org/10.1016/j.stueduc.2019.100812>
- Schroeders, U., & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. *European Journal of Psychological Assessment*, 26(4), 284–292. <https://doi.org/10.1027/1015-5759/a000038>
- Schwichow, M., Christoph, S., Boone, W. J., & Härtig, H. (2016). The impact of sub-skills and item content on students' skills with regard to the control-of-variables strategy. *International Journal of Science Education*, 38(2), 216–237. <https://doi.org/10.1080/09500693.2015.1137651>
- Schwichow, M., Osterhaus, C., & Edelsbrunner, P. A. (2020). The relation between the control-of-variables strategy and content knowledge in physics in secondary school. *Contemporary Educational Psychology*, 63, Article 101923. <https://doi.org/10.1016/j.cedpsych.2020.101923>
- Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, 58(2), 159–194. <https://doi.org/10.1007/BF02294572>
- Sheard, M. K., & Chambers, B. (2014). A case of technology-enhanced formative assessment and achievement in primary grammar: How is quality assurance of formative assessment assured. *Studies in Educational Evaluation*, 43, 14–23. <https://doi.org/10.1016/j.stueduc.2014.02.001>
- Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, 33, 1–19. <https://doi.org/10.1111/jcal.12172>
- Sternberg, R.J., Sternberg, K. (2012). *Cognitive Psychology*. Cengage Learning products. <https://doi.org/10.1039/ft9918702861>
- TIMSS, (1997). TIMSS Science Items: Released set for Population 2 (seventh and eighth grades). IEA TIMSS.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4–70. <https://doi.org/10.1177/109442810031002>
- Van Vo, D., & Csapó, B. (2020). Development of inductive reasoning in students across school grade levels. *Thinking Skills and Creativity*, 37(2020), Article 100699. <https://doi.org/10.1016/j.tsc.2020.100699>
- Van Vo, D., & Csapó, B. (2021aa). Exploring students' science motivation across grade levels and the role of inductive reasoning in science motivation. *European Journal of Psychology of Education*. <https://doi.org/10.1007/s10212-021-00568-8>
- Van Vo, D., & Csapó, B. (2021bb). Development of scientific reasoning test measuring control of variables strategy in physics for high school students: Evidence of validity and latent predictors of item difficulty. *International Journal of Science Education*, 1–21. <https://doi.org/10.1080/09500693.2021.1957515>
- Van Vo, D., & Csapó, B. (2022). Measuring inductive reasoning in school contexts: a review of instruments and predictors. *International Journal of Innovation and Learning*, 31(4), 506–525.
- Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, 44(3), 299–321. <https://doi.org/10.1080/00220272.2012.668938>
- Williamson, K. C., Williamson, V. M., & Hinze, S. R. (2017). Administering Spatial and Cognitive Instruments In-class and On-line: Are These Equivalent. *Journal of Science Education and Technology*, 26(1), 12–23. <https://doi.org/10.1007/s10956-016-9645-1>
- Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of Competencies in Educational Contexts* (pp. 83–110). Hogrefe & Huber.
- Wise, S. L., & Kuhfeld, M. R. (2021). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement*, 58(1), 130–149. <https://doi.org/10.1111/jedm.12275>
- Wood, K. E., Koenig, K., & Owens, L. (2018). Development of student abilities in control of variables at a two year college. *AURCO Journal*, 24, 164–179.