sustainability

MDPI

Article

Factors Associated with the Equivalence of the Scores of Computer-Based Test and Paper-and-Pencil Test: Presentation Type, Item Difficulty and Administration Order

Tzu-Hua Wang [1,*], Chien-Hui Kao [1] and Hsiang-Chun Chen [2]

[1] Department of Education and Learning Technology, National Tsing Hua University, Hsinchu City 300193, Taiwan; chienhuikao@gapp.nthu.edu.tw
[2] Department of Early Childhood Education, National Tsing Hua University, Hsinchu City 300193, Taiwan; hcchen@mail.nd.nthu.edu.tw
* Correspondence: tzuhuawang@gmail.com

Abstract: Since schools cannot use face-to-face tests to evaluate students' learning effectiveness during the COVID-19 pandemic, many schools implement computer-based tests (CBT) for this evaluation. From the perspective of Sustainable Development Goal 4, whether this type of test conversion affects students' performance in answering questions is an issue worthy of attention. However, studies have not yielded consistent findings on the equivalence of the scores of examinees' answering performance on computer-based tests (CBT) and paper-and-pencil tests (PPT) when taking the same multiple-choice tests. Some studies have revealed no significant differences, whereas others have exhibited significant differences between the two formats. This study adopted a counterbalanced experimental design to investigate the effects of test format, computerised presentation type, difficulty of item group, and administration order of item groups of different difficulty levels on examinees' answering performance. In this study, 381 primary school fifth graders in northern Taiwan completed an achievement test on the topic of Structure and Functions of Plants, which is part of the primary school Natural Science course. The achievement test included 16 multiple-choice items. After data collection and analysis, no significant differences in the answering performance of examinees were identified among the PPT, CBT with single-item presentation, and CBT with multiple-item presentation. However, after further analysis, the results indicated that the difficulty of item group and the administration order of item groups of different difficulty levels had significant influences on answering performance. The findings suggest that compared with a PPT, examinees exhibit better answering performance when taking multiple-choice tests in a CBT with multiple-item presentation.

Keywords: administration order; computer-based test; COVID-19 pandemic; item difficulty; multiple-choice test; paper-and-pencil test

Citation: Wang, T.-H.; Kao, C.-H.; Chen, H.-C. Factors Associated with the Equivalence of the Scores of Computer-Based Test and Paper-and-Pencil Test: Presentation Type, Item Difficulty and Administration Order. *Sustainability* **2021**, *13*, 9548. https://doi.org/10.3390/su13179548

Academic Editor: Clemens Mader

Received: 23 July 2021
Accepted: 20 August 2021
Published: 25 August 2021

1. Introduction

Since the year 2019, with the outbreak of the COVID-19 pandemic, many countries have curbed the spread of the virus by reducing crowd movement or close interaction, and temporarily closing certain places, such as educational institutions and public recreational places. Considering the equitable quality education and lifelong learning opportunities promoted by the United Nations on the Sustainable Development Goal 4, many schools applied online education to compensate for the learning impossibility at school during the pandemic and to mitigate the impact caused by school closures [1]. Since the information and communication technology (ICT) resources each student can access are different, which might affect the opportunities to learn and the fairness of evaluation, in addition to providing online courses to continue educational activities, schools have needed to modify the original evaluation standard and method to ensure the fairness of evaluation during the school closures [2,3]. As students are unable to take face-to-face paper-and-pencil tests,

take-home exams—one of the most common alternative assessment methods—are often delivered online in the form of CBT, allowing students to read on screen or in paper-based format at home [1,2]. In other words, CBT is a common alternative to evaluate students' learning effectiveness during the COVID-19 pandemic.

In addition to the COVID-19 pandemic era, CBT has gradually been applied to various assessment activities. The most important reason for this trend is that CBT can provide advantages such as varied test items, automatic scoring, and reduced error rates in real time. CBTs analyse and provide the required feedback immediately to facilitate recording and querying, and test items and shared item databases can be generated easily [4,5]. Because of the high efficiency of automatic scoring and analysis, teachers can also conduct formative assessment through CBT during the teaching process to immediately understand students' learning status and adjust their teaching methods accordingly. Further, examinees can also receive feedback from the test results and in turn cultivate their self-assessment ability, which enables CBT to play an active role in the teaching process [6]. In addition to transforming the original paper-and-pencil test (PPT) into computerised implementation, the real-time computing advantages of CBT have gradually been applied to develop the computer-based adaptive test based on item response theory to more accurately assess cognitive ability of examinees on standardised tests such as the TOEFL-iBT and GRE. The development of multimedia technology has also led to the introduction of multimedia-based test items, which are used to evaluate the abilities of examinees that cannot be evaluated by traditional tests [7].

Although CBTs have many advantages, their reliability and validity are still questioned by many researchers because they differ from the traditional answering process of PPTs, and test performance may be affected by the convenience of the answering process of the test and the examinees' personal computer experiences, which raises the question of differential validity between CBT and PPT [8]. This suggests that the scores of CBT and PPT may not have equivalence or interchangeability. Many organisations have proposed guidelines, such as ATP Computer-Based Testing Guidelines [9], International Guidelines on Computer-Based and Internet Delivered Testing [10], and The Standards for Educational and Psychological Testing [11].These guidelines emphasise that when interpreting CBT scores, any effects caused by computers must be eliminated or recalculated to avoid the influence of factors unrelated to evaluated abilities, including computer anxiety and computer experiences, on test scores. In addition, these guidelines also suggest that in the process of CBT, attention should be paid to the interchangeability of scores of different testing procedures or test forms administered in different formats. However, research has not provided consistent evidence of the equivalence and interchangeability of scores between CBT and PPT. For example, Leeson [12], Wang et al. [13], and Dadey [14] indicated that when the same test items are administered in PPT and CBT, the examinees may have different performance levels because of the differing presentations of test items. The factors influencing performance include screen size, font size, number of lines, interline spacing, whitespace, scrolling, item review, item presentation, resolution of graphics, and the presence of multiscreen, graphical, or complex displays. This indicates that when PPT and CBT are used to administer the same test items, they may not be measuring the same construct. The findings are consistent with Pommerich's [15] argument. Pommerich determined that when long reading texts or test items cannot be fully presented on the screen, users must use the mouse, which increases the complexity of the computer interface and affects answering performance. Leeson [12] also pointed out that only a single test item displayed on the screen may make the examinees answer in a rush and in turn lead to an increase in the error rate and have negative influences on their performance. On the other hand, if multiple items are displayed on the screen and examinees are allowed to preview all the items to be answered and provided with functions such as skip, scan, and build off previous item information on the screen, the "facilitating effect" can be achieved and it has a positive impact on answering performance [12]. Pommerich [16] further evaluated more than 3000 students in grades 11 and 12 to measure the differences in the item difficulty

index of the same test administered as a PPT and CBT. The test items included lengthy passages, and the topics included reading and science reasoning. The CBT was divided into paging and scrolling groups. The results revealed that in the reading test, the average scores of the PPT group were higher than those of the CBT group, but the difference was not significant. On the science reasoning test, the average scores of the CBT group were higher than those of the PPT group, but only the paging group achieved statistically significant differences, and the correlation coefficient among the difficulty of the CBT and PPT groups was greater than 0.9. Examinees' attitudes and familiarity with computers also influence their answering performance in CBT. Russell and Plati [17] indicated that if examinees are familiar with typing on a computer, and a CBT is used to evaluate their writing ability, the students may exhibit better performance than in a PPT. Wang et al. [18] also found that compared with PPT, examinees' attitudes regarding CBT were more positive, so they exhibited better answering performance. In addition, the subject is another important factor. Kingston [19] conducted a meta-analysis of 81 studies about the differences in scores in subjects such as mathematics, reading, English language arts, science, and social studies when tests were administered using CBT and PPT among American students in grades 1–12 from 1997 to 2007. The results showed that grade level had no influence on the results, but the subjects exhibited significant influences. When CBT was adopted for English language arts and social studies, the examinees had better answering performance, but when PPT was adopted in mathematics, answering performance was better. Besides, Hensley [20] also identified a significant difference between CBT and PPT in the test performance of 142 college students in the mathematics courses.

However, some studies have indicated that the scores of CBT and PPT have equivalence and interchangeability. For example, Wang et al. [13] and Logan [21] compared mathematics tests administered using CBT and PPT through a literature review and meta-analysis and found no statistically significant differences. Several studies also found no significant difference between CBT and PPT in examinees' answering performance in the reading comprehension tests [22], mathematics test [23], language learning test [24,25]. The findings of relevant studies are summarised in Table 1.

Table 1. Factors influencing examinees' answering performance in CBT and PPT.

Category	Factor	Explanation	Significant Influence	Study
Presentation factors	Display	Screen size, font size and style, resolution of graphics and screen, multiscreen, graphical or complex displays, line length, number of lines, interline spacing, whitespace, scrolling	Yes	Wang et al.(2007) Leeson (2006) Dadey (2018)
	Answering strategy	Reviewing and revising previous responses	Yes	Wang et al.(2007) Leeson (2006)
Content factors	Subjects	English language, arts, reading, social studies, mathematics	No	Kingston (2009)
		Reading comprehension	No	Pommerich (2007) Hosseini et al. (2014)
		English proficiency	No	Retnawati (2015) Khoshsima et al. (2019)
		Mathematics	No	Wang et al. (2007) Hamhuis et al. (2020)
		Science	No	Hamhuis et al. (2020)
		Mathematics	Yes	Hensley (2015)
		Science reasoning	Yes	Pommerich (2007)

Based on relevant research, the possible factors influencing examinees' answering performance in CBT and PPT can be investigated by dividing between the presentation

factors and the content factors. Presentation factors refer to the means used to present test information, including screen size, font size, resolution of graphics, the nature of the display (multiscreen, graphical, or complex), the amount of test information that can be presented on a screen at one time, and interactive assessment strategies. Content factors refer to the content of test items, including test subjects such as English, social studies, culture, mathematics, reading, and scientific reasoning. Table 1 shows that the influence of the presentation factors is larger than that of content factors. In traditional PPT, several content factors may affect performance outcomes, including item difficulty and the distribution of items at different difficulty levels. The sequence of items at different difficulty levels may affect examinees' test anxiety, and in turn it may affect their confidence, and ultimately having an impact on their performance outcomes [26–28]. However, studies (see Table 1) have not offered further investigation on the influence of these content factors. In addition, Leeson [12] stated that some early studies pointed out that in the CBT, whether the test items are displayed in single-item presentation or multiple-item presentation can influence an examinee's answering performance. However, Leeson also pointed out that the forms of test item design in these studies are all affected by the limitation of technologies. Therefore, it is necessary to further explore this factor and other factors that may affect the answering performance in CBT, so as to further understand the potential impacts of various CBT designs on the answering performance. The ways to address these effects should also be further investigated.

In addition, the PPT originally used at school were converted into CBT via the internet during the school closure caused by the COVID-19 pandemic [29]. From the perspective of Sustainable Development Goal 4, whether different test formats will affect students' performance in answering questions is an issue worthy of attention [29]. This is because whether students' performance under different test formats can be converted equally will cause fairness problems [8]. To address this research gap, in this study, a counterbalanced experimental design was used to explore how the answering performance of examinees differs when they answer multiple-choice items in CBT and PPT. This study also explored how the computerised presentation type (single-item presentation vs. multiple-item presentation), difficulty of item group, and administration order of item groups of different difficulty levels influence examinees' answering performance in PPT and CBT. Thus, the study addressed the following research question:

- What are the effects of test format (CBT and PPT), computerised presentation type, difficulty of item group, and administration order of item groups of different difficulty levels on students' answering performance in CBT and PPT?

2. Methodology

2.1. Participants

The participants were fifth-grade students from 16 classes of two primary schools in northern Taiwan, including 199 boys (average age: 11.1 years) and 182 girls (average age: 10.9 years) for a total of 381 students (average age: 11.0 years). In order to avoid the impact of participants' unfamiliarity with the computer, Internet, and CBT environment on the research results [30], all students had taken basic courses related to computers and the Internet. Before participating in this study, they had practised using the CBT environment of this study. After completing this study, each participant received stationery as a reward for joining the experiment.

2.2. Instruments

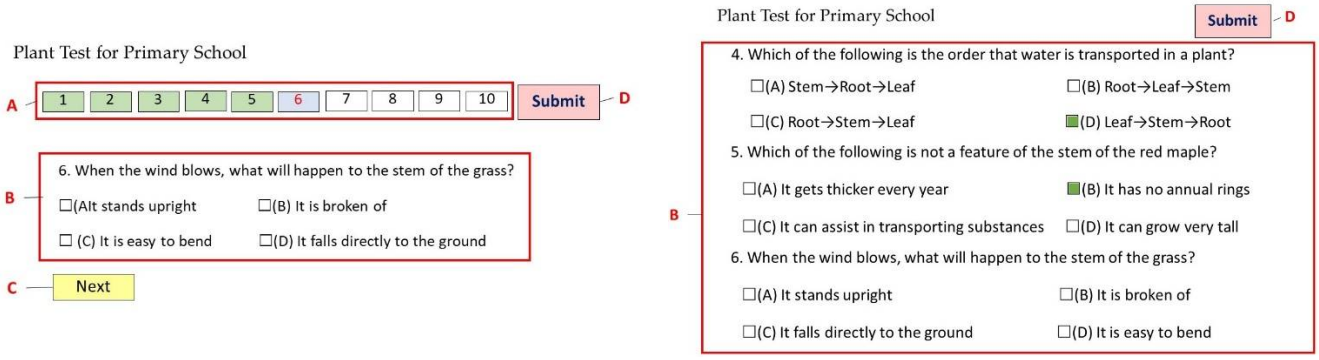
2.2.1. Achievement Test

Based on the Structure and Functions of Plants topic from the primary school fifth-grade Natural Science course, 20 multiple-choice items were designed. All 20 items were tested by 136 fifth graders who had learned about the topic, and the difficulty and discrimination indexes of each item were evaluated. The higher the difficulty index, the less difficult the item was; the lower the difficulty index, the more difficult the item was. After

the difficulty index of each item was evaluated, the eight simplest items were chosen to form the simple-item group (difficulty index mean = 0.772, SD = 0.099), and the eight most difficult items were chosen to form the difficult-item group (difficulty index mean = 0.348, SD = 0.076). There is a significant difference between the difficulty index mean of difficult-item group and simple-item group ($t(14) = 9.621$, $SE = 0.441$, $p < 0.001$). The achievement test adopted in this study consisted of the simple-item group and the difficult-item group. Both group items separately adopted the CBT and PPT test format. The simple-item and difficult-item groups were used to understand the effects of their administration order and item difficulty on students' answering performance in CBT and PPT.

2.2.2. Computer-Based Test Environment

The CBT in this study was presented in the Web browser. The computerised presentation types of the CBT were divided into single-item presentation (Figure 1a) and multiple-item presentation (Figure 1b), where Zone A is the item number area, showing the items already answered (green background), the items now being answered (blue background), and the items yet to be answered (white background). Zone B is the item and option presentation area, where the examinee can click on the box of an option to choose it as the correct answer; Zone C is pressed to present the next item after the answer is confirmed; and Zone D is used to submit all answers.



(**a**) (**b**)

Figure 1. Two computerised presentation types. In (**a**), Zone A is the item number area and there is only one item being answered in Zone B. Examinees can know their answering progression from Zone A. In (**b**), there are several items being answered in Zone B. Examinees can use mouse scrolling features to read other items.

2.3. Research Design

To address the research questions, this study investigated the effects of test format, computerised presentation type, difficulty of item group, and administration order of item groups of different difficulty levels on students' answering performance in the achievement test. Regarding test format, the achievement test was presented in the form of PPT and CBT. In the PPT, all items of the achievement test were presented on one test paper. The CBT had two computerised presentation types, namely, single-item presentation (CS) and multiple-item presentation (CM). The CS version presented one item at a time, and the examinee would press the 'NEXT' button after answering to present the next item. The CM version presented all the test items on the computer screen at a time. In terms of the difficulty of item group, the items of the achievement test were divided into the simple-item and difficult-item groups according to the difficulty index, with eight items in each group. Regarding the administration order of item groups of different difficulty levels, the test was divided into two types: the simple-item group followed by the difficult-item group, or the difficult-item group followed by the simple-item group. Finally, a counterbalanced experimental design was adopted to form eight treatments (see Appendix A). In the within-subject design, the factors of test format (PPT and CBT) and difficulty of item

group (simple-item and difficult-item) were adopted. In the between-subject design, the computerised presentation types (CS and CM) and administration order of item groups of different difficulty levels (the administration order of simple-item group and difficult-item group) were adopted. Taking one class as a unit, participants from 16 classes were randomly assigned to a treatment which comprised two classes of students.

2.4. Data Collection and Analysis

The data collected in this study were all quantitative data (i.e., examinees' correct answering rate on the achievement test in the eight treatment types). To address the research questions, this study adopted an independent sample *t*-test to perform descriptive analysis. In addition, this study adopted two-way ANOVA, taking test format, administration order of item groups of different difficulty levels, difficulty of item group, and computerised presentation type as independent variables and the correct answering rate on the achievement test as the dependent variable.

3. Results

3.1. Test Item Analysis

This study first compared the average correct answering rate of items on the achievement test under different test formats and item groups of different difficulty levels, as shown in Table 2. The results of the independent sample *t*-test reveal that the average correct answer rates of the PPT and CS as well as the PPT and CM did not exhibit significant differences ($t(30) = -0.220$, $SE = 0.089$, $p = 0.827$, $ES = 0.076$, $t(30) = -0.139$, $SE = 0.091$, $p = 0.891$, $ES = 0.051$). This result suggests that without considering the factors of difficulty of item group and administration order of item groups of different difficulty levels, the correct answering rates on the achievement test were not significantly different based on whether the test items were presented in a PPT and CS or in a PPT and CM.

Table 2. Descriptive analysis on the average correct answering rate of achievement test by different test formats and item groups of different difficulty levels.

Test Format	Simple-Item Group			Difficult-Item Group			Total		
	N	ACAR	SD	N	ACAR	SD	N	ACAR	SD
PPT	8	0.726	0.163	8	0.291	0.130	16	0.509	0.266
CS	8	0.709	0.163	8	0.348	0.130	16	0.528	0.235
PPT	8	0.726	0.166	8	0.299	0.114	16	0.530	0.276
CM	8	0.734	0.135	8	0.353	0.139	16	0.543	0.237

Note. PPT: paper-and-pencil test; CS: computerized single-item presentation; CM: computerized multiple-item presentation; N: number of items; ACAR: average correct answering rate; SD: standard deviation.

3.2. Analysis of Answering Performance of Simple-Item Group and Difficult-Item Group by Test Format and Administration Order of Item Groups of Different Difficulty Levels

This study tested the effects of test format and administration order of item groups of different difficulty levels on the answering performance of the achievement test. The following is an analysis of whether the effect of the independent variables on the dependent variables was different when the test items were presented in the form of a PPT and CS from when the test items were presented in the form of a PPT and CM. The effects of different test formats and different administration orders of item groups of different difficulty levels on examinees' answering performance in the simple-item group and the difficult-item group were analysed as follows.

3.2.1. Simple-Item Group

First, the simple-item group was examined using two-way ANOVA to determine the correct answering rate of all examinees under the different test formats and different administration orders of item groups of different difficulty levels (see Tables 3 and 4). For the PPT and CS as well as the PPT and CM, in the simple-item group, the main effect

of the test format was not significant ($F_{1,191}$ = 0.397, MSe = 0.049, p = 0.530, η^2 = 0.002; $F_{1,186}$ = 1.278, MSe = 0.035, p = 0.260, η^2 = 0.007), indicating that the test format had no significant influence on the correct answering rate. The main effect of administration order was also not significant ($F_{1,191}$ = 1.116, MSe = 0.049, p = 0.292, η^2 = 0.006; $F_{1,186}$ = 2.265, MSe = 0.035, p = 0.134, η^2 = 0.012), suggesting that the administration order had no significant influence on the correct answering rate. Moreover, the two variables did not exhibit a significant interaction effect ($F_{1,191}$ = 0.005, MSe = 0.049, p = 0.946, η^2 = 0.000, $F_{1,186}$ = 0.083, MSe = 0.035, p = 0.773, η^2 = 0.000).

Table 3. Descriptive analysis on the average correct answering rate of simple-item group in the achievement test under different test formats and different administration orders.

Test Format	Administration Order of Simple-Item Group						Total		
	First			Second					
	N	ACAR	SD	N	ACAR	SD	N	ACAR	SD
PPT	48	0.711	0.197	49	0.742	0.237	97	0.727	0.218
CS	49	0.689	0.224	49	0.725	0.227	98	0.707	0.225
PPT	42	0.789	0.145	49	0.740	0.187	91	0.762	0.170
CM	50	0.750	0.219	49	0.717	0.185	99	0.734	0.203

Note. PPT: paper-and-pencil test; CS: computerised single-item presentation; CM: computerised multiple-item presentation; First: simple-item group is administered in the first part of the achievement test; Second: simple-item group is administered in the second part of the achievement test; N: number of items; ACAR: average correct answering rate; SD: standard deviation.

Table 4. Summary table of two-way ANOVA.

		Source	*SS*	*df*	*MS*	*F*	*p*	η^2
SI	PPT&CS	Between						
		Test format	0.020	1	0.020	0.397	0.530	0.002
		Administration order	0.055	1	0.055	1.116	0.292	0.006
		Test format x Administration order	0.000	1	0.000	0.005	0.946	0.000
		Error	9.396	191	0.049			
	PPT&CM	Between						
		Test format	0.045	1	0.045	1.278	0.260	0.007
		Administration order	0.080	1	0.080	2.265	0.134	0.012
		Test format x Administration order	0.003	1	0.003	0.083	0.773	0.000
		Error	6.534	186	0.035			
DI	PPT&CS	Between						
		Test format	0.156	1	0.156	4.917	0.028 *	0.025
		Administration order	0.061	1	0.061	1.924	0.167	0.010
		Test format x Administration order	0.181	1	0.181	5.705	0.018 *	0.029
		Error	6.043	191	0.032			
	PPT&CM	Between						
		Test format	0.159	1	0.159	4.633	0.033 *	0.024
		Administration order	0.001	1	0.001	0.036	0.850	0.000
		Test format x Administration order	0.000	1	0.000	0.004	0.950	0.000
		Error	6.408	187	0.034			

Note. * $p < 0.05$; PPT: paper-and-pencil test; CS: computerised single-item presentation; CM: computerised multiple-item presentation; SI: simple-item group; DI: difficult-item group.

3.2.2. Difficult-Item Group

Next, we used two-way ANOVA to analyse the difficult-item group, examining the correct answering rate of all examinees under different test formats and different administration orders of item groups of different difficulty levels (see Tables 4 and 5). We

observed that in the PPT and CS, the main effect of test format was significant ($F_{1,191} = 4.917$, $MSe = 0.032$, $p < 0.05$, $\eta^2 = 0.025$), suggesting that test format had a significant influence on the correct answering rate. The correct answering rate in the CS was significantly higher than that in the PPT, but the effect of administration order was not significant ($F_{1,191} = 1.924$, $MSe = 0.032$, $p = 0.167$, $\eta^2 = 0.010$), suggesting that administration order had no significant influence on the correct answering rate. The two variables exhibited a significant interaction effect ($F_{1,191} = 5.705$, $MSe = 0.032$, $p < 0.05$, $\eta^2 = 0.029$). We conducted post hoc analyses and found that in the difficult-item group administered in PPT, the effect of administration order was significant; the correct answering rate when the difficult-item group was presented in the first part was significantly higher than when the difficult-item group was presented in the second part ($t(95) = 2.932$, $SE = 0.033$, $p = 0.004$, $ES = 0.288$). In the difficult-item group administered in CS, the effect of administration order was not significant; the correct answering rate when the difficult-item group was presented in the first part was not significantly different from when the difficult-item group was presented in the second part ($t(96) = -0.665$, $SE = 0.039$, $p = 0.514$, $ES = 0.067$). When the difficult-item group was presented in the first part, the effect of the test format was not significant; no significant difference was identified between the PPT and CS ($t(95) = 0.116$, $SE = 0.037$, $p = 0.908$, $ES = 0.012$). However, when the difficult-item group was presented in the second part, the effect of the test format was significant; the correct answering rate of the CS was significantly higher than that of the PPT ($t(96) = -3.393$, $SE = 0.035$, $p = 0.001$, $ES = 0.327$).

Table 5. Descriptive analysis on the average correct answering rate of difficult-item group in the achievement test under different test formats and different administration orders.

Test Format	Administration Order of Difficult-Item Group						Total		
	First			Second					
	N	ACAR	SD	N	ACAR	SD	N	ACAR	SD
PPT	48	0.339	0.177	49	0.242	0.145	97	0.290	0.168
CS	49	0.334	0.192	49	0.360	0.194	98	0.347	0.192
PPT	49	0.299	0.165	51	0.292	0.138	100	0.295	0.151
CM	49	0.355	0.214	42	0.351	0.218	91	0.353	0.215

Note. PPT: paper-and-pencil test; CS: computerised single-item presentation; CM: computerised multiple-item presentation; First: difficult-item group is administered in the first part of the achievement test; Second: difficult-item group is administered in the second part of the achievement test; N: number of items; ACAR: average correct answering rate; SD: standard deviation.

For the PPT and CM, the main effect of the test format was significant ($F_{1,187} = 4.633$, $MSe = 0.034$, $p < 0.05$, $\eta^2 = 0.024$), suggesting that test format had a significant influence on the correct answering rate; the correct answering rate in the CM was significantly higher than in the PPT, but the effect of the administration order was not significant ($F_{1,187} = 0.036$, $MSe = 0.034$, $p = 0.850$, $\eta^2 = 0.000$); the two variables did not exhibit a significant interaction effect ($F_{1,187} = 0.004$, $MSe = 0.034$, $p = 0.950$, $\eta^2 = 0.000$).

Based on the findings above, the different test formats (PPT, CS, and CM) and different administration orders of the simple-item group (first part and second part) had no significant influence on the correct answering rate in the achievement test. This study revealed that in the difficult-item group, the test format and administration order had significant influences on the correct answering rate of the achievement test: (1) PPT and CS—in the PPT, compared with when the difficult-item group is presented in the second part, the correct answering rate was significantly higher when the difficult-item group was presented in the first part; when the difficult-item group was presented in the second part, the correct answering rate in CS was significantly higher than that in the PPT; (2) PPT and CM—the correct answering rate of difficult-item group in the CM was significantly higher than that in the PPT; administration order had no significant influences on the correct answering rate. In other words, for the difficult-item group, the significant influences of different test formats (PPT, CS, CM) on the correct answering rate in the achievement test could be observed.

3.3. Analysis of Answering Performance of the Achievement Test by Difficulty of Item Group and Computerised Presentation Type

Another research purpose of this study was to examine the influences of the computerised presentation type (CS and CM) and the difficulty of item group (simple-item group, difficult-item group) on answering performance in the achievement test. This study analysed whether the effects of the independent variables on the dependent variables were different in different computerised presentation types.

Two-way ANOVA was used to examine the correct answering rate of all examinees in different computerised presentation types and item groups of different difficulty levels (see Tables 6 and 7). The results suggest that the main effect of computerised presentation type was not significant ($F_{1,187} = 0.931$, $MSe = 0.038$, $p = 0.336$, $\eta^2 = 0.005$). The computerised presentation type did not have a significant influence on the correct answering rate in the achievement test. However, the main effect of the difficulty of item group reached a significant level ($F_{1,187} = 263.023$, $MSe = 0.050$, $p < 0.01$, $\eta^2 = 0.584$), suggesting that the difficulty of item group had a significant influence on the correct answering rate on the achievement test, and the simple-item group was significantly higher than the difficult-item group; no significant interaction effect between the two variables was observed ($F_{1,187} = 0.327$, $MSe = 0.050$, $p = 0.568$, $\eta^2 = 0.002$).

Table 6. Descriptive analysis on the average correct answering rate of achievement test items under different computerised presentation type and different difficulty of item groups.

Difficulty of Item Group	Computerised Presentation Type						Total		
	CS			CM					
	N	ACAR	SD	N	ACAR	SD	N	ACAR	SD
SI	98	0.707	0.225	91	0.739	0.203	189	0.722	0
DI	98	0.347	0.192	91	0.353	0.215	189	0.350	0.203

Note. CS: computerised single-item presentation; CM: computerised multiple-item presentation; SI: simple-item group; DI: difficult-item group; N: number of items; ACAR: average correct answering rate; SD: standard deviation.

Table 7. Summary table of two-way ANOVA.

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
Between						
Computerised presentation type	0.035	1	0.035	0.931	0.336	0.005
Error	7.013	187	0.038			
Within						
Difficulty of item group	13.118	1	13.118	263.023	0.000 **	0.584
Difficulty of item group x Computerised presentation type	0.016	1	0.016	0.327	0.568	0.002
Error	9.327	187	0.050			

Note. ** $p < 0.01$.

On the basis of these findings, the computerised presentation type (CS and CM) demonstrated no significant influence on the correct answering rate in the achievement test; only the difficulty of item group exhibited a significant influence on the correct answering rate. The correct answering rate of the simple-item group was greater than that of the difficult-item group.

4. Concluding Remarks

The purpose of this study was to explore how the formats of multiple-choice tests influence the examinees' answering performance when they are administered as CBT and PPT. It is anticipated that several suggestions can be given for schools that are forced to change the test format for the evaluation of students' learning effectiveness, from PPT to CBT, during the COVID-19 pandemic. The conversion of test format might affect students' performance in answering questions, thereby affecting the accuracy and fairness

of evaluation on students' learning effectiveness [8]. This study analysed four factors: test format, computerised presentation type, difficulty of item group, and administration order of item groups of different difficulty levels. The four factors were investigated using a counterbalanced experimental design. Regarding test format, the findings showed that comparing PPT, CBT with single-item presentation, and CBT with multiple-item presentation revealed no significant difference in answering performance of the examinees; this is consistent with previous research results [12,13,16,21–23,25]. However, this study further examined the factors of difficulty of item group and administration order of item groups of different difficulty levels and revealed that the examinees' performance in the CBT and PPT was significantly different only for the difficult-item group. The CBT with multiple-item presentation (CM) was the most favourable in terms of examinee performance. This result can be explained by Leeson's viewpoint. Leeson points out that CBT adopts the multiple-item presentation method, which gives examinees a chance to preview test items and improve their answering performance. This is called the "facilitating effect" [12]. Regardless of whether the difficult-item group was arranged as the first or second part of the test, the performance of the examinees was better than that in the PPT. In the CBT with single-item presentation (CS), only when the difficult-item group was arranged as the second part of the test did the examinees perform better than in the PPT. For the difficult-item group, whether the CS or CM was used, the examinees' performance was better than that in the PPT, and the answering performance of examinees was not significantly different in the CS or CM. Based on the findings, the present study suggests that test administrators may consider adopting the form of CM on multiple-choice tests so that examinees can achieve better answering performance.

Additionally, the findings indicate that the difficulty of test items may influence the equivalence of examinees' answering performance in the CBT and PPT. This may explain the inconsistent findings on the equivalence of examinee's answering performance in the CBT and PPT in previous studies. The difficulty of test items can influence examinees' answering performance in CBT and PPT, as explained by cognitive load theory [31]. The efficiency of learners' cognitive operation can be influenced by cognitive load, and the sources of cognitive load include intrinsic cognitive load, extraneous cognitive load, and germane cognitive load [31,32]. In this study, cognitive load theory suggests that item difficulty influences intrinsic cognitive load, whereas the test format, computerised presentation type, and administration order of item groups of different difficulty levels influence the extraneous cognitive load and germane cognitive load of examinees during the testing process. In other words, if the same test implemented in the traditional PPT format is changed to the CBT format, it may affect extraneous cognitive load and germane cognitive load. Therefore, with a limited response time, to enable the examinees to perform at their highest ability and achieve a successful answering performance, we must pay attention to how cognitive load influences examinees' responses on the test. If the examinees' answering performance is influenced by cognitive load, then the test itself cannot effectively measure the real abilities of the examinees and achieve the test goals.

According to this concept, because the traditional exam is generally administered in PPT, if CBT is adopted, the extraneous cognitive load of the examinees is be increased because the test requires additional operation of mouse and keyboards as well as searching for the items, options, and answer area on the computer screen. If the examinees are not familiar with computer operation, the extraneous cognitive load is more severe. If the load is within the range of the examinees can bear, it will not influence their answering performance. However, if the test items themselves are difficult, meaning the intrinsic cognitive load of examinees is high, then the extraneous cognitive load, which increases due to the change to CBT, will cause the overall cognitive overload of examinees to influence their answering performance. This concept may explain the findings of this study; that is, the item difficulty will affect the equivalence of students' performance in the CBT and PPT, especially for the most difficult items. Additionally, this study finds that during the exam, whether the difficult-item group is presented by the CM or by the CS (and the

difficult items are arranged in the second part of the test), the examinees' performance is better than that in the PPT. This can be explained by the suggestions of Miller et al. [26] on test preparation, which indicate that the distribution of test items should be based on the principle of "from simple to difficult" to improve the performance of anxious examinees. In other words, it is beneficial to answering performance when examinees answer the simple-item group first and then the difficult-item group after. In addition, according to Mayer's [33] suggestions on multimedia learning environment design, when managing the cognitive load in a multimedia environment, the "segmenting principle" is helpful to learners. It helps provide learners with germane cognitive load and effectively manage their cognitive load [33,34]. In the present study, in both the CS and CM, only one or a few items were presented on the computer screen at a time, allowing the examinees to answer items in CBT more attentively than in PPT, as they were not being affected by other items in the test. This aligns with the "segmenting principle", helping examinees focus on answering the difficult-item group with its higher intrinsic cognitive load. In other words, for the difficult-item group, if the design of the answering environment can be oriented to reduce the extraneous cognitive load and increase the germane cognitive load, the answering performance may improve.

This study has some limitations. During the school closure caused by the COVID-19 pandemic, schools implemented "take-home exams" to prevent the epidemic and ensure the equitable quality education advocated by the Sustainable Development Goal 4. However, the technical issues, students' academic integrity, and testing environment may affect evaluation results. That is, if students do not follow the honest principles of traditional face-to-face paper-and-pencil test, encounter technology problems, or stay in an uncomfortable testing environment, the results of "take-home exams" have no reference value for the evaluation of students' learning effectiveness [2,3,29,35]. This study is mainly to understand the equivalence of scores of different test formats. In order to avoid the aforementioned factors that affect students' performance in answering questions will impact on the results of the study, the CBT and PPT are not implemented in the way of "take-home exams" in this study. As a consequence, findings might have some limitation in the application in the context of COVID-19 pandemic. In addition, because this study mainly concerned fifth graders of primary school and focused on investigating the multiple-choice items for the specific topic of primary school Nature Science course, the research results cannot be directly extended to examinees of other ages, other subjects, or other types of test items. It is suggested that further research should be conducted with other age groups and subjects. This study also suggests follow-up investigation on other types of test items that can be implemented in both CBT and PPT, such as filling blanks, short answer questions, and essay questions to further explore the difference of examinees' answering performance on the same test in the CBT and PPT. In addition, the CBT is a new test format for fifth graders in primary schools. Although before this study, all the examinees had taken basic courses related to computers and the Internet and practised operating in a CBT environment, they may still have experienced a novelty effect [7]. Their self-efficacy and perceptions towards CBT may have influenced the research results [17,18,30]. Therefore, this study suggests that future researchers should seek to increase the experience of examinees participating in the CBT and provide them with rich CBT experience before conducting the research, thereby reducing the influences of the novelty effect and computer operation familiarity on the research results.

This study also revealed that, generally, the examinees' answering performance was not significantly different in the CBT and PPT. However, when answering more difficult test items, compared with the PPT, examinees exhibited better answering performance in the CBT. In the CBT with multiple-item presentation, the administration order of difficult and simple items has no significant effect. However, in the CBT with single-item presentation, it is suggested that the difficult items be placed in the later part of the test, so that examinees can have better answering performance. In other words, when the same test items were implemented in the CBT or PPT, examinees' answering performance in the CBT may be

better than that in the PPT. This finding suggests that a CBT can encourage examinees to leverage their abilities to answer multiple-choice items. It also indicates that the real abilities of examinees may be measured more effectively in CBT. In the COVID-19 pandemic, as distance education is becoming an important strategy to maintain educational equity, accordingly, remote CBT would become an important way for the evaluation of learning effectiveness. Based on the findings of this research, if the teacher wants to use the CBT format, and the content is multiple-choice items with standard answers, it will be a suitable type whether the CBT is administered in the way of multiple-item presentation or single-item presentation. Special attention should be paid to the item difficulty and their administration order in the test. If there is a significant difference in the difficulty of items in the test, it is suggested to administer the CBT by using multiple-item presentation, that is, presenting all items for students to answer at one time, and students are free to choose which one to answer first. If the CBT is presented in a single-item presentation format, students have to answer items in a sequence arranged by the computer without autonomy in choosing the order of answering. In this respect, the difficult items should be placed in the later part of the test. In addition, this study suggests referring to Mayer's [33] suggestions in follow-up studies to make full use of computer functions to design various methods of presenting test items and various answering strategies to help reduce or assist examinees to manage their cognitive load during the answering process. Through this approach, a CBT can become a more effective and appropriate test format in the era of e-Learning.

Author Contributions: Conceptualization, T.-H.W. and H.-C.C.; methodology, T.-H.W. and H.-C.C.; software, T.-H.W.; validation, T.-H.W.; formal analysis, T.-H.W. and H.-C.C.; investigation, T.-H.W. and C.-H.K.; resources, T.-H.W. and C.-H.K.; data curation, T.-H.W. and C.-H.K.; writing—original draft preparation, T.-H.W., H.-C.C. and C.-H.K.; writing—review and editing, T.-H.W.; visualization, T.-H.W. and C.-H.K.; supervision, T.-H.W.; project administration, T.-H.W.; funding acquisition, T.-H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science and Technology in Taiwan, grant number 106-2511-S-007-003-MY3 and 109-2511-H-007-007-MY3.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Acknowledgments: The authors deeply appreciate the Ministry of Science and Technology in Taiwan for the financial support and encouragement under Grant No. 106-2511-S-007-003-MY3 and 109-2511-H-007-007-MY3.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Counterbalanced experimental design.

Treatment	Administration Order in the Achievement Test	
	First Part	Second Part
1	Simple-item group in PPT	Difficult-item group in CS
2	Simple-item group in PPT	Difficult-item group in CM
3	Difficult-item group in PPT	Simple-item group in CS
4	Difficult-item group in PPT	Simple-item group in CM
5	Simple-item group in CS	Difficult-item group in PPT
6	Simple-item group in CM	Difficult-item group in PPT
7	Difficult-item group in CS	Simple-item group in PPT
8	Difficult-item group in CM	Simple-item group in PPT

Note. PPT: paper-and-pencil test; CS: computerised single-item presentation; CM: computerised multiple-item presentation.

References

1. Elsalem, L.; Al-Azzam, N.; Jum'ah, A.A.; Obeidat, N.; Sindiani, A.M.; Kheirallah, K.A. Stress and behavioral changes with remote E-exams during the Covid-19 pandemic: A cross-sectional study among undergraduates of medical sciences. *Ann. Med. Surg.* **2020**, *60*, 271–279. [CrossRef] [PubMed]
2. Gamage, K.A.; Silva, E.K.D.; Gunawardhana, N. Online delivery and assessment during COVID-19: Safeguarding academic integrity. *Educ. Sci.* **2020**, *10*, 301. [CrossRef]
3. Guangul, F.M.; Suhail, A.H.; Khalit, M.I.; Khidhir, B.A. Challenges of remote assessment in higher education in the context of COVID-19: A case study of Middle East College. *Educ. Assess. Eval. Account.* **2020**, *32*, 519–535. [CrossRef]
4. Parshall, C.G.; Spray, J.A.; Kalohn, J.C.; Davey, T. *Practical Considerations in Computer-Based Testing*; Springer: New York, NY, USA, 2020. Available online: https://link.springer.com/book/10.1007%2F978-1-4613-0083-0 (accessed on 18 July 2021).
5. Wang, T.H. Developing a web-based assessment system for evaluating examinee's understanding of the procedure of scientific experiments. *Eurasia J. Math. Sci. Technol. Educ* **2018**, *14*, 1791–1801. [CrossRef]
6. Wang, T.H. Developing web-based assessment strategies for facilitating junior high school students to perform self-regulated learning in an e-learning environment. *Comput. Educ.* **2011**, *57*, 1801–1812. [CrossRef]
7. Wang, T.H.; Kao, C.H.; Dai, Y.L. Developing a web-based multimedia assessment system for facilitating science laboratory instruction. *J. Comput. Assist. Learn.* **2019**, *35*, 529–539. [CrossRef]
8. Zou, X.L.; Ou, L. EFL reading test on mobile versus on paper: A study from metacognitive strategy use to test-media impacts. *Educ. Assess. Eval. Acc.* **2020**, *32*, 373–394. [CrossRef]
9. Association of Test Publishers. ATP Computer-Based Testing Guidelines. 2002. Available online: http://www.testpublishers.org (accessed on 18 July 2021).
10. International Test Commission (ITC). International Guidelines on Computer-Based and Internet Delivered Testing. 2004. Available online: http://www.intestcom.org (accessed on 18 July 2021).
11. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing*; American Educational Research Association: Washington, DC, USA, 2014.
12. Leeson, H.V. The mode effect: A literature review of human and technological issues in computerized testing. *Int. J. Test.* **2006**, *6*, 1–24. [CrossRef]
13. Wang, S.; Jiao, H.; Young, M.J.; Brooks, T.; Olson, J. A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educ. Psychol. Meas.* **2007**, *67*, 219–238. [CrossRef]
14. Dadey, N.; Lyons, S.; DePascale, C. The comparability of scores from different digital devices: A literature review and synthesis with recommendations for practice. *Appl. Meas. Educ.* **2018**, *31*, 30–50. [CrossRef]
15. Pommerich, M. Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *J. Tech. Learn. Assess.* **2004**, *2*. Available online: https://ejournals.bc.edu/index.php/jtla/article/view/1666 (accessed on 18 July 2021).
16. Pommerich, M. The effect of using item parameters calibrated from paper administrations in computer adaptive test administrations. *J. Tech. Learn. Assess.* **2007**, *5*. Available online: https://ejournals.bc.edu/index.php/jtla/article/view/1646 (accessed on 18 July 2021).
17. Russell, M.; Plati, T. Does it matter with what I write? Comparing performance on paper, computer and portable writing devices. *Curr. Issues Educ.* **2002**, *5*. Available online: http://cie.ed.asu.edu/volume5/number4/ (accessed on 18 July 2021).
18. Wang, S.; Young, M.J.; Brooks, T.E. *Administration Mode Comparability Study for Stanford Diagnostic Reading and Mathematics Tests (Research Report)*; Harcourt Assessment: San Antonio, TX, USA, 2004.

19. Kingston, N.M. Comparability of computer-and-paper-administered multiple-choice tests for K-12 populations: A synthesis. *Appl. Meas. Educ.* **2009**, *22*, 22–37. [CrossRef]
20. Hensley, K.K. Examining the Effects of Paper-Based and Computer-Based Modes of Assessment of Mathematics Curriculum-Based Measurement. Ph.D. Thesis, University of Iowa, Iowa, IA, USA, 2015. [CrossRef]
21. Logan, T. The influence of test mode and visuospatial ability on mathematics assessment performance. *Math. Educ. Res. J.* **2015**, *27*, 423–441. [CrossRef]
22. Hosseini, M.; Abidin, M.J.Z.; Baghdarnia, M. Comparability of test results of computer based tests (CBT) and paper and pencil tests (PPT) among English language learners in Iran. *Pro. Sco. Behav. Sci.* **2014**, *98*, 659–667. [CrossRef]
23. Hamhuis, E.; Glas, C.; Meelissen, M. Tablet assessment in primary education: Are there performance differences between TIMSS'paper-and-pencil test and tablet test among Dutch grade-four students? *Br. J. Educ. Technol.* **2020**, *51*, 2340–2358. [CrossRef]
24. Retnawati, H. The comparison of accuracy scores on the paper and pencil testing vs. computer-based testing. *Turk. Online J. Educ. Technol.-TOJET* **2015**, *14*, 135–142.
25. Khoshsima, H.; Hashemi Toroujeni, S.M.; Thompson, N.; Reza Ebrahimi, M. Computer-based (CBT) vs. paper-based (PBT) testing: Mode effect, relationship between computer familiarity, attitudes, aversion and mode preference with CBT test scores in an Asian private EFL context. *Teach. Engl. Technol.* **2019**, *19*, 86–101.
26. Miller, M.D.; Linn, R.L.; Gronlund, N.E. *Measurement and Assessment in Teaching*, 11th ed.; Pearson: New York, NY, USA, 2012.
27. Ollennu, S.N.N.; Etsey, Y.K.A. The impact of item position in multiple-choice test on student performance at the basic education certificate examination (BECE) level. *Univers. J. Educ. Res.* **2015**, *3*, 718–723. [CrossRef]
28. Nie, Y.; Lau, S.; Liau, A.K. Role of academic self-efficacy in moderating the relation between task importance and test anxiety. *Learn. Individ. Differ.* **2011**, *21*, 736–741. [CrossRef]
29. Camara, W. Never let a crisis go to waste: Large-scale assessment and the response to COVID-19. *Educ. Meas.* **2020**, *39*, 10–18. [CrossRef]
30. Nardi, A.; Ranieri, M. Comparing paper-based and electronic multiple-choice examinations with personal devices: Impact on students' performance, self-efficacy and satisfaction. *Br. J. Educ. Technol.* **2019**, *50*, 1495–1506. [CrossRef]
31. Sweller, J.; Ayres, P.; Kalyuga, S. Measuring cognitive load. In *Cognitive Load Theory*; Springer: New York, NY, USA, 2011; pp. 71–85. Available online: https://link.springer.com/chapter/10.1007/978-1-4419-8126-4_6 (accessed on 18 July 2021).
32. Sweller, J. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educ. Psychol. Rev.* **2010**, *22*, 123–138. [CrossRef]
33. Mayer, R.E. Using multimedia for e-Learning. *J. Comput. Assist. Learn.* **2017**, *33*, 403–423. [CrossRef]
34. Singh, A.M.; Marcus, N.; Ayres, P. The transient information effect: Investigating the impact of segmentation on spoken and written text. *Appl. Cogn. Psychol.* **2012**, *26*, 848–853. [CrossRef]
35. Raje, S.; Stitzel, S. Strategies for effective assessments while ensuring academic integrity in general chemistry courses during COVID-19. *J. Chem. Educ.* **2020**, *97*, 3436–3440. [CrossRef]