Test Delivery Medium Matters: Cognitive Effort Exertion on Assessment

Tuncer AKBAY [1], Lokman AKBAY [2], Osman EROL [3]

To Cite: Akbay, T., Akbay, L. & Erol, O. (2021). Test delivery medium matters: Cognitive effort exertion on assesment. *Malaysian Online Journal of Educational Technology, 9*(2), 76-85. http://dx.doi.org/10.52380/mojet.2021.9.2.273

[1] Asst. Prof.
tuncerakbay@mehmetakif.edu.tr, Burdur Mehmet Akif Ersoy University, Turkey
ORCID: 0000-0003-3938-1026

[2] Asst. Prof.
lokmanakbay@iuc.edu.tr, Istanbul University-Cerrahpasa, Turkey
ORCID: 0000-0003-4026-5241

[3] Asst. Prof.
oerol@mehmetakif.edu.tr, Burdur Mehmet Akif Ersoy University, Turkey
ORCID: 0000-0002-9920-5211

ABSTRACT

Integration of e-learning and computerized assessments into many levels of educational programs has been increasing as digital technology progresses. Due to a handful of prominent advantages of computer-based-testing (CBT), a rapid transition in test administration mode from paper-based-testing (PBT) to CBT has emerged. Recently, many national and international testing agencies have been offering an electronic version of some low- and high-stake tests along with their paper versions. In this study, we aim to examine test administration mode effect from a standpoint of cognitive effort exertion. To this end, the results of this experimental study suggest that the cognitive effort exertion rates of CBT and PBT examinees are different. More specifically, the study results suggest empirical evidence that examinees exert higher cognitive effort in a CBT in comparison to its PBT counterpart.

Keywords: *Computer-based-testing; paper-based-testing; cognitive effort; cognitive process, CBT, media in education*

Article History:
Received: 30 Nov. 2020
Received in revised form: 19 Mar. 2021
Accepted: 31 Mar. 2021

INTRODUCTION

Technological progress and the digital revolution allowed educators to benefit from modern computer technology in instructional design. The advantage of computer technology in education gave rise to the integration of e-learning and computerized assessments into many levels of educational programs. There is a notable transition in the test administration mode or test delivery method from paper-based-test (PBT) administration to computer-based-test (CBT) administration. Immediate feedback due to reduced "time-lag" between the administration and score [*]reporting; ease of scoring, analyzing, and reporting on the assessment; greater test security; flexibility in testing schedule; reduced cost; ability to include multimedia test items; enabling customized/tailored tests; and allow of measuring the response time may be counted among the main advantages of CBTs over PBTs (Bennet, 2001; Parshall, Spray, Kalohn, & Davey, 2002). The CBT may also be beneficial for learners with disabilities as they may offer further accessibility features such as text-to-speech support, built-in video, and large-print (Kim & Huynh, 2010).

Given the advantages of CBTs, the shift in test administration has moved beyond the school walls to high-stake and large-scale assessments. Due to the visible advantages of CBT, many national and

[*] This study was piloted with a different sample using different measurement method and the preliminary findings were presented at the 27th international Conference on Educational Sciences, Antalya, Turkey. The presentation was titled 'Cost of performing a test: Traditional vs. computerized'.

international testing agencies started to offer an electronic version of some low- and high-stake tests along with the paper versions. Examples include but not limited to state-wide end-of-course (EOC) English exam (Kim & Huynh, 2010), CBT options for end-of-course, end-of-grade, and high school graduation examinations in many states of the USA (Luecht & Sireci, 2011), and a comprehensive foreign language exam, namely YDS, offered by the Assessment, Selection and Placement Center in Tukey. However, along with the increased interest in the CBT versions of PBTs, there are increased concerns on the accuracy of the ability estimation due to the mismatch between the test delivery methods of tests (Russell & Haney, 2000). The difference in the test performance due to delivery methods of the same test is referred to as the test administration mode effect (Clariana & Wallace, 2002; Pomerich, 2004). Standards for educational and psychological testing set forth by AERA, APA, and NCME highlight the importance of presenting empirical evidence indicating that no one is disadvantaged by the new test mode (AERA, APA, & NCME, 1999).

To address the concerns about the test administration mode effects on academic achievement, several true-experimental and quasi-experimental studies have been conducted in various subject areas (e.g., Chua & Don, 2013; Clariana, & Wallace, 2002; DeAngelis, 2000; Hosseini, Abidin, & Baghdarnia, 2014; Garas & Hassan, 2018; Kim & Huynh, 2010; Kingston, 2009; Maguire, Smith, Brailler, & Palm, 2010; Mason, Patry, & Bernstein, 2001; Piaw, 2012; Russell, 1999; Pomplun, Frey, & Becker, 2002). These studies have mainly focused on either technological issues or participant characteristics to investigate the test administration mode effect. Screen resolution, interface features, and font size are the typical technological factors investigated in many studies (e.g., Bridgeman, Lennon, & Jackenthal, 2003; Vispoel, 2000). Likewise, gender, age, ethnicity, technology familiarity, and test mode preference are frequently examined variables about examinees (e.g., Clariana & Wallace, 2002; Gallagher, Bridgeman, & Cahalan, 2000; Garas & Hassan, 2018; Hensley, 2015). On top of the variation on academic achievement, some researchers have also studied the impact of test administration modes on test duration.

The above-mentioned research put forward contradictory results on both achievement and test duration so that advantages of one or the other test mode, if any, on achievement and test duration are not clear. We argue that considering the impact of cognitive effort exertion may provide useful information to explain the conditions where one or the other test administration mode may be more beneficial. Thus, in this study, our primary purpose is to reveal the impact of test administration mode on momentarily cognitive effort exertion that may or may not adversely impact test achievement. Furthermore, the secondary purpose of this research is to try to pull educators' and testing institutions' attention to equity and fairness of the test scores obtained from the CBT and PBT versions of the same test when both forms are available to examinees. Thus, this experimental study seeks the answer to the following research question:

"Is there any difference in examinees' momentarily cognitive effort exertion levels while taking PBT and CBT versions of the same test?"

Some brief explanation on the key terms such as the computer-based-tests and cognitive effort may be needed here. Traditional PBTs can be transformed into their electronic versions via an adaptive or non-adaptive algorithm. A computerized test incorporating an adaptive algorithm is referred to as a computerized adaptive test (CAT). CAT algorithm allows selection and administration of test items from an item pool to develop tailored tests for examinees considering their ability levels that are estimated from the responses of examinees to the items administered up to that point in test administration (Akbay & Kaplan, 2017; Wang & Shin, 2010). Contrary to the CAT, a linear transformation of PBTs to computer-based-tests (CBTs) does not employ an adaptive algorithm in item selection so that items, test procedure, and scoring method are the same in both PBT and CBT versions of a test (Wang & Shin, 2010). After making this distinction between the two types of PBT transformation, it should be noted here that, the current study only focuses on the linear CBTs (i.e., non-adaptive computer-based-tests) and PBTs.

In the cognitive psychology literature, cognitive effort refers to fractional working memory capacity and attentional resources partitioned among the simultaneously running tasks requiring cognitive processing (Olive & Barbier, 2017; Piolat, Olive, & Kellogg, 2005). Resources for working memory and attention are referred to as cognitive resources, which have a limited capacity (Rendell, 2010). Thus, one should balance the limited cognitive capacity and successful completion of given cognitive tasks. Therefore, successful completion of a cognitive task is possible only when there are sufficient cognitive resources to perform the task. On the contrary, a cognitive task requiring more cognitive resources than available is doomed to failure (McDowd, 2007).

Some studies on instructional design (e.g., Clark, 2002) argue that computer-based and paper-based versions of any teaching component will yield equal outcomes. Using a food and truck metaphor, Clark argues that the effect of the media (i.e., delivery method) used for teaching is no greater than the impact of a delivery-truck on the nutrition of our food (2002, 1994, 1983). According to Clark and his followers, media cannot go beyond being a medium such that media used in any component of education including measurement and assessment should not have any effect on student performance. The ongoing media-method debate between Clark and his followers advocating the effect of teaching methods rather than the media on the learning process and the opponents emphasizing the impact of media on the learning process is referred to as 'media-method endless debate' (Akyol & Cagıltay, 2007; Yazıcı & Kultur, 2020; Yang, Wang, & Chiu, 2014).

The media-method endless debate has been so much focused that, as highlighted by Yang et al. (2014), the effect of media on cognitive processes in technology-supported learning environments has not been sufficiently examined. Thus, in this study, we try to clarify why one should expect media to impact examinees' test performance. Referring back to Clark's metaphor, even if the vehicles used in the delivery of our food do not affect food nutrition; we claim that they may affect the delivery cost. Specifically for educational testing situations, we expect that elevation of examinees' momentarily cognitive effort rates due to test administration mode may result in depletion of their cognitive resources earlier. Especially in cases where a test requires either intensive momentarily cognitive efforts or longer testing duration, examinees may underperform due to lack of sufficient cognitive resources to complete the test.

RESEARCH METHOD

As stated earlier, the study purpose is to investigate test administration mode effect on examines' level of momentarily cognitive effort exertion. More specifically, we aim to understand the possible effect of test delivery mode on performance through the effect of the media tools on cognitive processes. Therefore, based on the theory of limited cognitive resources toward completion of a given cognitive task (Rendell, 2010), we aim to examine the examinees' momentarily cognitive effort exertion levels under PBT and CBT cases. Any statistically significant difference in cognitive resource depletion levels between the two test administration modes may be used to explain the variation in achievement under various testing conditions. To this end, an experimental study was conducted.

Participants

A total of 58 volunteers majoring in teaching English as a second language program in a public university have participated in the study. After informing them about the aim and procedure of the study, they signed and returned informed consent forms. The sample consisted of 65% female, 35% male participants. The participants were randomly assigned either to the control group (taking the PBT version of the test) or to the experimental group (taking the CBT version of the test).

Variables and the measurement tools

Due to its convenience, the authors of this manuscript chose to scrutinize *Foreign Language Exam (FLE)* test to disclose the impact of the test delivery method on momentarily cognitive effort exertion. FLE is a high-stakes nation-wide standardized foreign language test and its results are used in application procedures for various types of community service jobs, academic promotions, and graduate school programs. FLE has been administered as a PBT for decades and its linear CBT version, referred to as e-FLE, has been administered since 2014. Both forms of the test contain 80 multiple-choice items and offered multiple times within a calendar year. Thus, examinees have the opportunity to take either FLE or e-FLE and the obtained scores are valid for a few years for many purposes.

For this study, we administered the first 42 items of a publicly available e-FLE trial test. We preferred to work with the first 42 out of 80 items to keep our participants out of boredom, which may influence their performance throughout the test. The administered test consisted of 42 multiple-choice items; where (1) 26 fill-in-the-blank items where 10 out of 26 items were derived from two passages and the rests were stand-alone items; (2) six items required examinees to choose the most accurate translation of given sentences; and (3) the remaining 10 items required examinees to choose the expression or the subordinate clause that best completes the main clause. We designed our PBT version of e-FLE trial test such that it was comparable to its CBT counterpart. The item stems and response options were left-aligned and 1.5 spaced. Also, we inserted additional double-space between items. Throughout the booklet, Times New Roman font and 12 punto font-size were employed. Only two passages in the test were written in **bold**, as was the case in the CBT version of the same test.

The CBT version of the test was presented in a 21.5-inch monitor where the screen resolution was set to 1920×1080. Stem and response options of items were presented together on the screen. In other words, working on an item did not require examinees to scroll the screen up and down. To respond to an item, they clicked on a radio button next to response options. Examinees moved across items by clicking left and right arrows located on both edges of the screen. The review button located at the top of the screen such that it allowed examinees to jump directly to the interested test item. To do so, examinees would click the *review* button and then select an item number from the appearing list. Once the examinees completed the test, they were able to save the answers and finalize the test by clicking on the *end of the exam* button located at the top-right corner of the screen. As can be understood from the description, the interface of the used program was easy to navigate and the screen allowed examinees to see entire items without scrolling.

Measurement of cognitive effort

As defined earlier, cognitive effort refers to fractional working memory capacity and attentional resources partitioned among the simultaneously running tasks that require cognitive processing (Piolat et al., 2005; Olive & Barbier, 2017). According to general capacity theory, a more cognitive effort requiring tasks require more attention (Abernethy, 1988). In cases where a task requires more cognitive resources than available, the task is doomed to failure (McDowd, 2007). One of the well-known techniques for measuring cognitive effort is the dual-task technique, which relies on the assumption that, given a primary and a secondary task simultaneously, these two tasks compete for the greater share of limited working memory capacity (Engle, 2002).

In the dual-task technique, cognitive effort used in the primary task is predicted by the performance in the secondary task. Specifically, performance in the secondary task tends to decrease with an increase in cognitive demands of the primary task (Olive, Alves, & Castro, 2009). In dual-task, the primary task is the one that researcher the aims to measure cognitive effort used for it. Due to its attention-demanding nature, for the purpose of momentarily measuring cognitive effort exertion, reaction to given auditory stimuli (i.e., "beeps") was regarded as the secondary task in several studies (i.e., Akbay, 2018; Olive & Barbier, 2017; Olive, et al., 2009; Piolat, et al., 2005). For the purpose of our study, test-taking is the primary task while reaction to auditory stimuli (i.e., beep sound) is regarded as the secondary task. A software program referred to as ScriptKell (Piolat, Olive, Roussey, Thunin, & Ziegler, 1999) was developed and used to measure cognitive effort allocated to the cognitive activities as a means of measuring participants' reaction to auditory stimuli in real-time. Thus, in this study, we used the ScriptKell program for the same purpose.

Procedure

First of all, a single-session achievement test was administered to PBT and CBT groups to which participants were randomly assigned. This session lasted about 90 minutes along with the time spent for preparation and instructions for the testing. Examinees in the CBT group were provided with the desktop computers to take the electronic version of the test, while the examinees in the PBT group were seated on a desk with the PBT version of the test (i.e., booklet), optical mark recognition sheet, pencil, and eraser. Regardless of the test administration modes, all participants required to wear a headset, through which the examinees were provided with the auditory stimuli for the secondary task. Moreover, all examinees were provided with a desktop computer running the SkriptKell program to record examinees reaction times to auditory stimuli. Examinees were asked to right-*click* on the mouse as soon as they hear the auditory stimuli so that SkriptKell could record their reaction times (i.e., time lag between the stimuli and clicking the mouse). Because examinees' reaction times would ultimately be used to estimate their momentarily cognitive effort exertion, to avoid extraneous influence such as a delay due to switching windows, CBT group members were provided with separate computers to take the e-FLE test.

Just before the administration of these tests, the researchers took the examinees' baseline reaction time to the secondary task. The baseline is basically the reaction time to auditory stimulus in the absence of the primary task of test-taking. To define examinees baseline reaction times (RTs), 14 random auditory stimuli were presented to examinees. These stimuli were presented randomly with a time interval of 10 to 20 seconds. In this process, the very first four RTs were treated as warm-up activities and were not counted toward the calculation of the baseline for RT. Then, the baseline score for each examinee was computed by taking the mean of the remaining 10 RTs. After taking the baseline, the auditory stimuli were also presented during the administration of the foreign language test. Throughout the test administration, auditory stimuli were randomly presented within every 90 to 120 seconds and participants were asked to continue reacting to the auditory stimuli as they work on the primary task of test-taking. Using a stopwatch, we also kept track of the participants' primary task duration. At the end of the test, examinees responses, RT scores, and task durations were recorded.

To derive the cognitive effort exertion, differences in RT scores ($\Delta RT = RT_{Test} - RT_{Baseline}$) were calculated. These differences in RT scores (when the secondary task running alone vs when the primary and secondary task running simultaneously) are attributed to the cognitive effort used for the primary task by the examinees. Thus, these ΔRTs were considered as the estimates of momentarily used cognitive effort for completing the FLE test. In dual-task technique, the larger ΔRT stands for the larger the momentarily cognitive effort exertion. To quantify the total cognitive effort demand of a task, momentarily used cognitive effort should be multiplied by the task duration (Christensen-Szalanski, 1980; Cooper-Martin, 1994). Accordingly, to determine participants' total cognitive effort exertion, we multiplied ΔRT scores measured in milliseconds by the test durations measured in minutes.

In the end, we measured achievement and cognitive effort exertion for both PBT and CBT groups. For the data analysis purposes, because our ultimate goal was to compare the groups' cognitive effort exertion levels, we conducted a statistical test on the group mean differences and the results are presented in the next section.

FINDINGS

Prior to hypothesis testing, the dataset was screened for outliers in terms of both achievement and cognitive effort exertion. Boxplots flagged five cases (2 in CBT and 3 in PBT) as significant outliers so that they were removed from the dataset. The descriptive statistics of the data set are presented in *Table 1.* Although the difference in the achievement levels of the two groups of examinees is not the main interest of this study, we compared it to check whether our random groups are significantly different in terms of achievement/ability levels. Because in cases where the achievement levels of the two groups are not equal, in comparison to the cognitive effort exertion levels, achievement levels should be controlled by assigning achievement scores as covariate.

Table 1. *Descriptive statistics*

Variables	Test Mode	n	Mean	SD	Skewness		Kurtosis	
					Statistic	S. Error	Statistic	S. Error
Achievement	PPT	26	21.29	5.41	-0.245	0.456	0.281	0.887
	CBT	27	21.20	5.60	-0.597	0.448	-0.110	0.872
Cognitive Effort	PPT	26	8416.83	9295.73	0.823	0.456	0.351	0.887
	CBT	27	19601.73	12111.46	-0.275	0.448	-0.506	0.872

Note: N = sample size; Mean = mean scores; SD = standard deviation; and S. Error = standard error.

After relevant assumptions were checked, an independent samples t-test was conducted to see whether test mode affects the achievement. Normality assumption checked through the skewness and kurtosis values (*see Table 1*) first, both values were between the critic values of -1.5 and +1.5 (Tabachnick & Fidell, L. S. (2013).

Additionally, normality was also checked through Shapiro-Wilk test and it was confirmed that the data met the normality assumption (W_{CBT} = 0.959, *p* = .355; W_{PBT} = 0.981, *p* = .887). Then, Levene's Test of Equality of Variances confirmed that the variances in the test performance score in CBT and PBT groups are equal *F* = 0.183, *p* = .671. As shown in *Table 2*, the independent samples t-test yielded a non-significant achievement score differences between the CBT and PBT conditions (*t* = -0.056, *p* = .956).

Table 2. *Independent samples t-tests*

Variables	Levene's test		Independent samples t-test			Effect size
	F-statistic	p-value	t-statistic	df	p-value	Measures
Achievement	0.183	.671	-0.056	51	.956	
Cognitive Effort	1.707	.197	3.761	51	.000	*d* = 1.04; η^2 = .21

Note: df = degrees of freedom; *d* = Cohen's *d*; and η^2 = eta-squared.

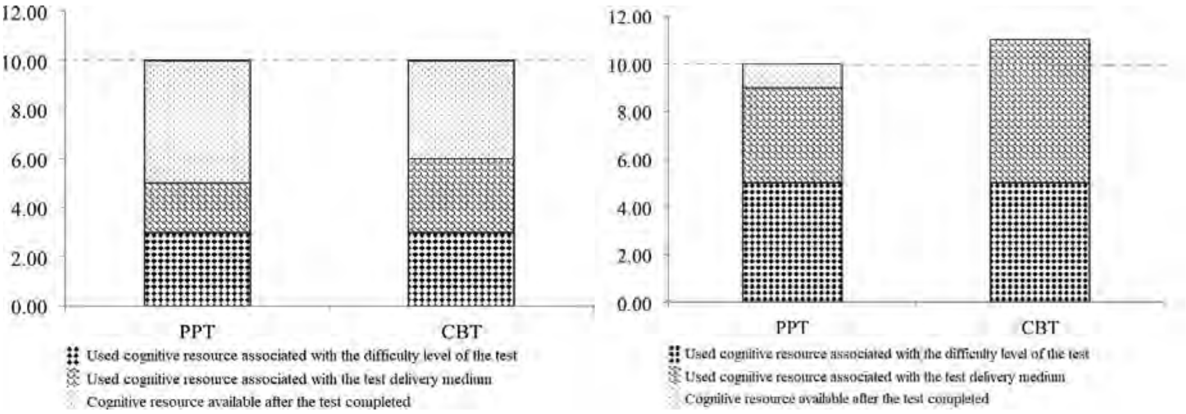
Because the results of the independent samples t-test yielded a non-significant result in testing the achievement levels, we proceeded with an independent samples t-test to see whether test mode affects the cognitive effort exertion. Again, before conducting the t-test, the normality assumption was checked through the skewness and kurtosis values (*see Table 1*); which were within the boundaries of -2 and +2. The normality was further checked with Shapiro-Wilk test and it was confirmed that the data met the normality assumption (W_{CBT} = 0.982, *p* = .902; W_{PBT} = 0.933, *p* = .093). Then, conducting Levene's Test of Equality of Variances, we checked the homogeneity of variance assumption. The test result confirmed that the variances in the test performance score in CBT and PBT groups were equal (*F* = 1.707, *p* = .197). Finally, we conducted an independent samples t-test to test the significance of the difference in cognitive effort exertion levels of the two groups. As given in *Table 2*, the test yielded a significant difference between the CBT and PBT groups (*t* = 3.761, *p* < .001). Furthermore, we computed the effect size measures (i.e., *d* = 1.04, η^2 = .21), which indicated a large effect for the test mode. From these results, we conclude that the test mode has a large effect on the cognitive effort levels in test-taking.

DISCUSSION AND CONCLUSION

The current study revealed non-significant test administration mode effect on the test scores of CBT and PBT versions of FLE. It means that the examinees taking either the CBT or PBT version of FLE had similar achievement scores from the test. This finding supported some of the findings of the other previously published research findings (Garas, & Hassan, 2018; Hosseini et al., 2014; Ita et al., 2015; Karay et al., 2015; Khoshsima & Hashemi Toroujeni, 2017; Kim & Huynh, 2010; Kingston, 2009; Piaw, 2012; Mason et al., 2001) and contradicted to the findings of some other research findings (Choi, Kim, & Boo 2003; Chua & Don, 2013; Clariana, & Wallace, 2002; Hosseini et al., 2014; Maguire et al., 2010). It should be reminded here that, as mentioned by Hensley (2015), many researchers claimed an achievement difference between PBT and CBT attributed this difference to either technological issues or participant characteristics. Equal test scores that the current study revealed might be explicated by the several facts: (1) by random examinee assignment the study design ruled out the possible impact of extraneous variables such as examinee characteristics, (2) the participants of the current study may be considered as digital natives so that the technology familiarity and the interface factor should not interfered with the test results, and (3) possible impact of the screen resolution could be ruled out due to improvement in computer displays.

Moreover, a significant difference between the CBT and PBT groups in cognitive effort exertion was reported in this study. Specifically, 21% of the variance in cognitive effort exertion is accounted for test delivery method (i.e., η^2 = .21). Hereby, no matter what the underlying causes are (i.e., participant characteristics and technological issues, etc.), this study discloses that the cognitive effort depletion levels of the CBTs and PBTs are different. More specifically, this study provides empirical evidence that examinees exert more cognitive effort when the test mode is CBT rather than PBT. Considering the fact that the cognitive resources are limited (Akbay, 2018; Muraven, Tice, & Baumeister, 1998; Muraven & Baumeister, 2000), the variation in the cognitive effort use, which may be referred to as cognitive cost, between CBTs and PBTs may affect the testing performance as the cognitive demand on testing (i.e., intrinsic load due to test difficulty and extrinsic load due to test mode) increases.

Because the cognitive resources to meet the cognitive cost are limited, as the test requires extended time to be completed, the cognitive resource depletion rate may eventually affect test performance as examinees start to run out of cognitive resources. Moreover, people subconsciously control their own behaviors based on the expenditure of limited inner resource to maximize their best interest in the long-term (Muraven & Baumeister, 2000). Therefore, as examinees' cognitive levels decrease, to save the remaining cognitive resources for subsequent tasks, their performance may not be as high as otherwise, their performance would be. Pommerich (2004) addresses two points that may support our argument here; (1) one should expect unequal relative performances across test modes and (2) intuitively, one may expect greater difference across test modes as the test gets more complex (i.e., difficult). Our study results may elucidate Pommerich's intuitive expectations by arguing that if one test mode requires more cognitive effort than another, in long run, this difference may affect examinees' test performance.



(a) Cognitive resource depletion in an easy test (b) Cognitive resource depletion in a difficult test

Figure 1. *Two hypothetical conditions depicting cognitive resource depletion levels*

The graphs given in *Figure* 1 may help us to illustrate our claim here. The graphs in the figure display cognitive effort depletion across examinees taking the same test through two different delivery methods. Given the examinees ability levels are equal and available cognitive resources are 10 units, as displayed in the first graph (on the left), we may expect equal achievement for these groups when the test items are relatively easy and the time required to complete the tests is short. Notice that the levels of cognitive resource depletion across the groups due to test mode may be different, however, this difference would not impact achievement as long as the examinees have a good portion of cognitive resources. Similarly, another hypothetical situation is depicted in the second graph (on the right) where the test is relatively more difficult and the test duration is longer. In this case, elevated cognitive resource depletion may be observed because of increased cognitive cost due to test difficulty and test administration mode. In such cases, when one or the other test mode group may underperform as examinees run out of the cognitive resources.

Finally, based on the observed results of this study, we may conclude that the CBT version of a test potentially results in the overspending of cognitive resources in comparison to its PBT counterpart. Then, when both versions of a test is offered, as is the case in our example of the FLE test, examinees taking the CBT version may be disadvantaged due to increased cognitive cost. In other words, examinees taking the PBT version of a test might allocate a greater share of their available cognitive resources to the given tasks rather than spending for the testing media. Therefore, in order not to put some examinees in a disadvantaged position, educational researchers, assessment specialists, and policy-makers must take great care to determine the factors that cause examinees to spend more cognitive resource in CBTs and take actions to minimize their adverse impact. To serve this purpose, we may design and conduct an experimental study to ascertain the impact of related factors that may be responsible for the overload in CBT yielding larger cognitive resource depletion. Lastly, this study specifically focused on the impact of media on testing. However, given the increasing interest in technology use and online teaching, in every component of education, the impact of digital media on students' cognitive cost must be carefully considered.

REFERENCES

Abernethy, B. (1988). Dual-task methodology and motor skills research: Some applications and methodological constraints. *Journal of Human Movement Studies*, *14*, 101–122.

Akbay, L., & Kaplan, M. (2017). Transition to multidimensional and cognitive diagnosis adaptive testing: An overview of cat. *The Online Journal of New Horizons in Education*, *7(1)*, 206–214.

Akbay, T. (2018). *Investigating the effects of individual interest and goal-orientation on ordinary and worthy performance* (Unpublished doctoral dissertation). Middle East Technical University, Ankara, Turkey.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bennett, R. E. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archive*, *9(5)*, 1–23.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Re- search*, *31(1)*, 51–60.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, *16(3)*, 191-205.

Choi, I. C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, *20(3)*, 295–320.

Chua, Y. P., & Don, Z. M. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. *Computers in Human Behavior*, *29(5)*, 1889–1895.

Clariana, R., & Wallace, P. (2002). Key factors associated with the test mode effect. *British Journal of Educational Technology*, *33(5)*, 593-602.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of educational research*, *53(4)*, 445–459.

Clark, R. E. (1994). Media will never influence learning. *Educational technology research and development*, *42(2)*, 21–29.

Cooper-Martin, E. (1994). Measures of cognitive effort. *Marketing Letters*, *5(1)*, 43–56.

Christensen-Szalanski, J. J. (1980). A further examination of the selection of problem- solving strategies: The effects of deadlines and analytic aptitudes. *Organizational Behavior & Human Performance*, *25(1)*, 107–122.

DeAngelis, S. (2000). Equivalency of computer-based and paper-and-pencil testing. *Journal of Allied Health*, *29*, 161–164.

Engle, R.W. (2002). Working memory capacity as executive attention. *Society*, *11(1)*, 19–23.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2000). The effect of computer-based tests on racial/ethnic, gender, and language groups. *ETS Research Report Series*, *2000(1)*, 117.

Garas, S., & Hassan, M. (2018). Student performance on computer-based tests versus paper-based tests in introductory financial accounting: UAE evidence. *Academy of Accounting and Financial Studies Journal*, *22(2)*, 1-14.

Goldberg, A. L., & Pedulla, J. J. (2002). Performance differences according to test mode and computer familiarity on a practice graduate record exam. *Educational and Psychological Measurement*, *62(6)*, 1053–1067.

Hensley, K. K. (2015). *Examining the effects of paper-based and computer-based modes of assessment on mathematics curriculum-based measurement* (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

Hosseini, M., Abidin, M. J. Z., & Baghdarnia, M. (2014). Comparability of test results of computer based tests (CBT) and paper and pencil tests (PPT) among English language learners in Iran. *Procedia - Social and Behavioral Sciences*, *98*, 659-667.

Ita, M. E., Kecskemety, K. M., Ashley, K. E., & Morin, B. C. (2015). Comparing student performance on computer-based vs. paper-based tests in a first-year engineering course. *The ASEE Computers in Education (CoED) Journal*, *6(1)*, 2–13.

Karay, Y., Schauber, S. K., Stosch, C., & Schuttpelz-Brauns, K. (2015). Computer versus paper - Does it make any difference in test performance? *Teaching and Learning in Medicine*, *27(1)*, 57-62.

Khoshsima, H., Morteza, S., Toroujeni, H., & Tefl, M. A. I. (2017). Comparability of computer-based testing and paper-based testing: Testing mode effect, testing mode order, computer attitudes and testing mode preference. *International Journal of Computer (IJC)*, *24(1)*, 80-99.

Kim, D. H., & Huynh, H. (2010). Equivalence of paper-and-pencil and online ad- ministration modes of the statewide English test for students with and without disabilities. *Educational Assessment*, *15(2)*, 107-121.

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

Luecht, R. M., & Sireci, S. G. (2011). *A Review of Models for Computer-Based Testing* (Research Report 2011-12). New York, NY: College Board.

Maguire, K. A., Smith, D. A., Student, G., & Palm, L. J. (2010). Computer-based testing: A comparison of computer-based and paper-and-pencil assessment. *Academy of Educational Leadership Journal*, *14(4)*, 117-138.

Mason, B. J., Patry, M., & Bernstein, D. J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. *Journal of Educational Computing Research*, *24(1)*, 29–40.

McDowd, J. M. (2007). An overview of attention: Behavior and brain. *Journal of Neurologic Physical Therapy*, *31(3)*, 98–103.

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*, *126(2)*, 247–259.

Muraven, M., Tice, D. M., & Baumeister, R. F. (1998). Self-control as a limited re- source: Regulatory depletion patterns. *Journal of personality and social psychology*, *74(3)*, 774-789.

Olive, T., Alves, R. A., & Castro, S. L. (2009). Cognitive processes in writing during pause and execution periods. *European Journal of Cognitive Psychology*, *21(5)*, 758-785.

Olive, T., & Barbier, M. L. (2017). Processing time and cognitive effort of longhand note taking when reading and summarizing a structured or linear text. *Written Communication*, *34(2)*, 224–246.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer.

Piaw, C. Y. (2012). Replacing paper-based testing with computer-based testing in assessment: Are we doing wrong? *Procedia - Social and Behavioral Sciences*, *64*, 655-664.

Piolat, A., Barbier, M. L., & Roussey, J. Y. (2008). Fluency and cognitive effort during first- and second-language notetaking and writing by undergraduate students. *European Psychologist*, *13(2)*, 114-125.

Piolat, A., Olive, T., & Kellogg, R. T. (2005). Cognitive effort during note taking. *Applied Cognitive Psychology*, *19(3)*, 291–312.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment*, *2(6)*, 1-45.

Pomplun, M., Frey, S., & Becker, D. F. (2002). The score equivalence of paper- and pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement*, *62(2)*, 337–354.

Rendell, M. A. (2010). *Cognitive effort in contextual interference and implicit motor learning* (Unpublished doctoral dissertation). Victoria University, Melbourne, Australia.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*, *7(20)*, 1–47.

Russell, M., & Haney, W. (2000). The gap between testing and technology in schools. *Education Policy Analysis Archives*, *8(19)*, 1–10.

Tabachnick, B. G., & Fidell, L. S. (2013). Using multivariate statistics (Sixth edition). United States: Pearson Education.

Vispoel, W. P. (2000). Computerized versus paper-and-pencil assessment of self- concept: Score comparability and respondent preferences. *Measurement and Evaluation in Counseling and Development*, *33(3)*, 130-143.

Wang, H., & Shin, C. D. (2010). Comparability of computerized adaptive and paper- pencil tests. *Test, Measurement and Research Service Bulletin*, *13*, 1–7.

Yazıcı, C., & Kultur, C. (2020). Medya mı yontem mi bitmeyen tartısma. In Cagiltay, K. & Goktas, Y. (Eds.), *Ogretim teknolojilerinin temelleri: Teoriler, arastirmalar, Egilimler* (pp. 123-142). Ankara: Pegem Akademi.

Yang, K.T., Wang, T.H., & Chiu, M.H. 2014. How technology fosters learning: Inspiration from the Media Debate, *Creative Education, 5,* 1086-1090.