Eurasian Journal of Educational Research
www.ejer.com.tr

Comparing Paper-Pencil and Computer-Based Tests: A Meta-Analysis Study in The Sample of Turkey

Funda NALBANTOGLU YILMAZ[1]

ARTICLE INFO

Article History:
Received: 25 Jan. 2020
Received in revised form: 14 Feb. 2021
Accepted: 12 Mar. 2021
DOI: 10.14689/ejer.2021.93.13

Keywords
computer-based test, paper-pencil test, effect size, meta-analysis

ABSTRACT

Purpose: With improvements in computer technologies and test implementations in the computer environment, when advantageous points of computer-based test implementations are considered, it is inevitable to compare psychometric characteristics of paper-and-pencil tests and computer-based tests and students' success. In computer-based tests, individuals' familiarity with computers and competency in using computers, conditions may show diversity depending on the country or region. Therefore, this study aims to investigate the effects of mean differences between PP and CBT using meta-analysis concerning the studies, including samples from Turkey and conducted between 1993 and 2020.

Research Methods: In this meta-analysis, 37 findings were included. Cohen's d was used as the effect size. And also, in this study, concerning the equivalence of the PP and CBT forms, it was investigated whether mean effect sizes differ or not according to variables like type of computerized, education level, and subject matter. In this direction, ANOVA and Q values were used.

Findings: As a result of the meta-analysis conducted, the general effect size was 0.042. In this direction, it was found that the difference in test implementation methods (paper-and-pencil, computer-based) was negligible.

Implications for Research and Practice: Results suggest that CBT can be an acceptable alternative to traditional pencil and paper tests. In this way, results obtained are expected to lead to educational policies and measurement implementations in the future.

[1] Nevşehir Hacı Bektaş Veli University, TURKEY, e-mail: fundan@nevsehir.edu.tr, ORCID: orcid.org/0000-0002-3228-4605

Introduction

Improvements in computer technology increase and even make it necessary to use information technologies in many points of our lives. Computers are used prevalently in both daily routine and also in every point of education quite often. In education, computer technologies are commonly used in identifying students, keeping data related to students, and at the same time, in the educational field and measurement and evaluation processes. The involvement of computer technologies in our lives to that extent makes computer-based testing implementations increase. Computer-based tests (CBT) can be applied in two ways as an implementation of paper-and-pencil tests in a computer environment (CBT-P) or computer-adaptive tests (CAT). The major difference between CAT and linear computerized tests is that length of the test questions individuals face and order of the questions can change depending on the ability of the individuals.

Together with improvements in information technologies and computer use in the world, it can be seen that CBT and/or CAT implementations increase. For instance, in many tests like TOEFL (Test of English as a Foreign Language), GRE (The Graduate Record Examination), GMAT (Graduate Management Admission Test), computer-adaptive test implementations are used. In Turkey, large-scale test implementations are conducted mostly in the form of the PP. In recent years, examinations like e-YDS (Foreign Language Proficiency Exam), e-ALES (Academic Personnel and Post-Graduate Education Entrance Exam) and Private Motor Vehicle Drivers' License Exam are implemented both on a paper-and-pencil basis and CBT in Turkey. However, I should note that computer-based tests, especially computer-adaptive tests remain limited in Turkey.

There are advantages of computerized assessment implementations in generating question bank, rapid scoring, quick-reporting of results, decrease in duration of exam, test security, being able to make assessment implementations in varied and demanded times, with individualized testing implementation, individuals' facing with questions at their own level of ability and providing measurement implementations for visually impaired individuals (Hambleton, Zaal, & Pieters, 1991). When those advantages are considered, it is expected to increase in implementation. Increment of computer-based and computer adaptive test methods in large-scale test implementations seem important because tests conducted once or twice a year lead to important decisions for individuals and the sake of solutions to problems in central examinations (Cikrikci-Demirtasli, 1999). With improvements in computer technologies and test implementations in the computer environment, when advantageous points of computer adaptive test implementations are considered, it is inevitable to compare psychometric characteristics of paper-and-pencil tests (PP) and computer-based tests and students' success. For example, do scores obtained via two methods show difference? Are ability estimates obtained from both implementations similar? Nowadays, studies comparing PP and CBT gradually increase. Combining and interpreting the knowledge obtained from studies conducted would provide valuable information to authorities and people who deal with psychometrics.

In the literature, many studies are comparing computer-based test implementations and in paper-and-pencil format. Scrutinizing the results of this study is thought to contribute to computer adaptive tests and PP implementations. In this direction, in this study, studies conducted on these subjects in Turkey are tried to be combined with the help of some statistical data and summarized systematically. For this purpose, meta-analysis, which is quantitatively synthesizing different study results, was used. Meta-analysis is a statistical method that is used to integrate, synthesize and interpret experimental findings obtained in individual studies (Wolf, 1986). In the meta-analysis, it is aimed to quantitatively combine different research results, which were conducted independently on any subject matter.

There are meta-analyses in literature that compare success and general ability. in examinations made on a paper-and-pencil basis and computer-based examinations from different years, different ages, different cultures, different grades and on different course/subjects. From these studies, Bergstrom (1992) conducted a meta-analysis study using 20 results obtained from eight studies and compared ability parameters obtained from CAT and PP between 1977 and 1992. Mead and Drasgow (1993) made a synthesis with studies published between 1977 and 1996 and comparing paper-and-pencil and computer-based implementations of cognitive ability tests of young adults. Finger and Ones (1999) synthesized studies examining whether the computerized form of the Minnesota Multiphasic Personality Inventory is psychometrically equal or not. Kim (1999) conducted a meta-analysis of 51 studies comparing PP and CBT or computer-adaptive tests between 1976 and 1996. Goldberg, Russell, and Cook (2003) conducted a meta-analysis between 1992–2002 focused on the comparison between K-12 students writing with computers and paper-pencil with 26 studies. In the study, significant mean effect sizes in favor of computers were found for the quantity of writing and quality of writing. Wang et al. (2007) conducted a meta-analysis of 38 findings in 14 studies written in the English language, which include paper-and-pencil and computerized implementations of mathematic courses of 12th Grade between 1980 and 2005. Wang et al. (2008) made a synthesis of the results of 42 independent research studies in the English language, including computer-based and paper-and-pencil implementations between the years of 1980 and 2005. Kingston (2009) synthesized 81 study results comparing computer-based and paper-and-pencil multiple-choice tests between 1997 and 2007 in the USA. Aybek et al. (2014) conducted a meta-analysis with 35 findings from nine studies that compare student success in PP and CBT implemented between 1999 and 2012 in Turkey and out of Turkey. The effect size was negligible in the literature between varied years, culture and subjects all PP-CBT meta-analysis comparisons.

The difference of this study from other meta-analysis studies is that it embodies studies that were conducted with samples from Turkey and conducted between 1993 and 2020. When the use of technology and familiarity with computers effects in education are taken into account based on the country and even based on districts, it is regarded as significant that this meta-analysis study comparing score or ability in exams conducted in CBT or PP formats must contain studies conducted with samples from Turkey. Because in computer-based tests, individuals' familiarity with computers, competency in using computers, conditions may show diversity depending on the country or region. In this direction, this study is the first meta-analytic study focusing on Turkey. In addition, the other meta-analysis includes only three studies in the Turkey sample. Also, some of the meta-analysis studies carried out includes a limitation of the subject matter. For instance, synthesis studies were conducted on studies with subjects of mathematics (Wang et al., 2007), reading skills (Wang et al., 2008), Minnesota Multiphasic Personality Inventory (Finger & Ones, 1999). Additionally, there is time-wise diversity between meta-analysis studies, including PP and CBT comparisons and this study. Meta-analysis studies in the literature include studies between 1974 and 2012. However, when computer literacy, familiarity, and competency, opportunity, school competency, improvements in computer technologies are considered, recent studies gain importance.

The study aims to determine the effects of mean differences between PP and CBT by using meta-analysis concerning the studies, including samples from Turkey and conducted between 1993 and 2020. In this research, answers were sought for the questions below.

1. What is the mean effect size between PP and CBT?
2. Does the effect size vary by the type of computerized method?
3. Does the effect size vary by education level?
4. Does the effect size vary by subject matter?

In this way, study results are expected to contribute to the field and to measurement implementations of large-scale testing programs, which are conducted frequently and lead to important decisions in the lives of individuals in Turkey about the way they are implemented. Moreover, it has also become important how measurement and evaluation should be performed in distance education during the pandemic experienced in the world. At this stage, online exams or computer-based test's similarity/reliability according to paper-pencil tests often come to the agenda.

Method

Research Design

This study was a meta-analysis study containing studies comparing measurement implementations conducted on computer-based and paper-and-pencil forms. The reason why a meta-analytical method was preferred in this study was that studies on this subject matter increase in literature, and in this direction, their contribution of statistically combining and interpreting the findings of independent and different studies. Moreover, the results obtained were expected to contribute to measurement implementations and test developers of the 21st century.

Identification of Studies

In determining the studies within the context of the research study, published and unpublished post-graduate and doctoral dissertations, published scientific research articles and papers are utilized. In this direction, through the Council of Higher Education, National Thesis Center, related Turkish and English dissertations were searched. Articles and papers were searched using a database like Google Academic, ULAKBIM Turkish National Databases-Journal-Park and ERIC. Congress and symposium proceeding books were searched. References of accessed studies were searched and by this way, new studies on the subject matter were also reached. Search was operated separately in Turkish and English languages. In searching, keywords like computer-based test, computerized test, computerized adaptive test, paper and pencil tests, computer-based and paper-and-pencil tests were used. The first study on this subject in Turkey was in 1993. Thus, all databases were searched between 1993 and 2020 (on February 05, 2020, and final search on May 12, 2020).

Criteria for Including Studies

1. Including obtained scores, student score/ability in measurement implementations of computer-based and paper-and-pencil test forms,

2. Including sufficient statistical detail (e.g., mean and standard deviation) for measuring the size of an effect,

3. Including parametric statistical methods,

4. Containing evidence of reliability and validity,

5. Conducted on students in Turkey,

6. Studies conducted between 1993 and 2020 were considered.

In some of the studies accessed, it was seen that subject matters like a comparison of computer-based linear and individualized test versions, developing computer adaptive test software, including only computer-adaptive tests, investigation of student opinions about paper-and-pencil and computer-based examinations, investigating student attitudes about computer-based examinations were dealt. Thus, these studies which did not contain PP and CBT comparison were excluded from this study. In some studies, insufficient information about results of the comparison, contain only correlation value, and also does not contain mean and standard deviation for the effect size calculation, these studies were also excluded from this study. Some studies did not meet normality assumptions, and parametric statistics were not used; these studies were also excluded from this study. Because the effect size used in this research is not robust to the violations of the normality assumption (Marfo & Okyere, 2019; Sun & Cheung, 2020). Studies that provide the normality assumption from the correlational studies and which contained mean and standard deviation values were included in the analysis.

After examining each study accessed, inclusion criteria were used to choose articles to be used in the research study. Accordingly, 37 findings from 21 separate studies were included in the meta-analysis.

Coding Procedure

Studies obtained from the literature review were investigated individually and by taking inclusion criteria into account, studies that enter into the research were determined. These studies were investigated and coded one by one. While coding the studies, descriptive characteristics like writer, year, the title of publication, type of research (e.g., article and dissertation), type of computerized, number of student-sample, educational level, course/subject matter, design of this study, statistical data of comparison results were used.

Calculating Effect Sizes (ES)

After the application of choosing criteria, the effect size was calculated with reference to coded study findings. "Effect size is simply a way of quantifying the size of the difference between two groups" (Coe, 2002, p.1). In addition, effect size provides a common metric to compare the direction and strength of the relationship between variables in this study (Berben, Sereika, & Engberg, 2012). Effect size is described as a statistical statement of the magnitude of the relationship between two variables or the magnitude of the difference between groups in terms of some interests. Depending on the aim and nature of the research, different effect sizes can be calculated. However, effect sizes like correlation coefficient r and Cohen's d may lead to very different results/interpretations about the same data (Falchikov & Boud, 1989; McGrath & Meyer, 2006). Therefore, it is mostly not appropriate to perform meta-analysis by combining effect sizes of relationship or experimental data between r and d. In combining/comparing effect sizes, it must be ensured whether it is related to the same results/learning outcomes (Coe, 2002). Also, just as stated by Borenstein, Hedges, Higgins, and Rothstein (2009), it should be questioned whether it is logical or not to include different effect sizes into the same analysis.

In the meta-analysis, it is more useful to user effect size when the purpose is to determine a relationship and it is better to use d effect size when the purpose is to determine the effect of an intervention (McGrath & Meyer, 2006). Because the aim of this study is to examine the mean difference between CBT and PP, the standardized mean difference (e.g., Cohen's d) was used as the effect size (ES = CBT-PP). Moreover, this meta-analysis includes between-participants design studies (n=12) and within-participants design studies (n=25). We used the ES of each study based on Cohen's d. In the literature, ES's from different research designs can be combined only if ES's from each design estimate the same treatment effect (Borenstein, Hedges, Higgins & Rothstein, 2009; Morris & DeShon, 2002). As recommended by Morris and DeShon (2002), we first calculated the ES of each study. Second, we transformed each ES into a common metric for comparison. Morris and DeShon (2002) recommend using the equation dBP= dWP $\sqrt{(2.(1-p)}$, where dBP is the ES for between-participants design, dWP is the ES for within-participants design, and ρ is the correlation. Moreover, we examined research design as a moderator effect. The studies' research design (between-participants design and within-participants design) was not a significant moderator ($Q_{(1)} = 0.223$, $p > .05$). Thus, there were no significant differences between the effect sizes for between and within participants' designs. To interpret the effect sizes, the following criteria were used. Cohen (1988) classified effect sizes are as "negligible, less than 0.2", "small effect, 0.2", "medium effect, 0.5", and "large effect, 0.8".

Statistical Independence

Sometimes, there might be more than one result in a study. Some criteria were used about whether these results would be used as separate study results or not. After carefully investigating studies, if the comparison of results reported in a study belongs to different student groups or different courses/subjects, it is stated that this is not dependent. It is accepted that such study findings are obtained from independent studies.

Fixed and Random Effects Models

In calculating the effect size, there are two methods as fixed and random. It is thought that the random effect model is more appropriate within the scope of research data because of differences like computer management systems, ages of students, course-subject in which the implementation is conducted and grade level. As far as heterogeneity is concerned, a random model has been proposed (Hedges & Olkin, 1985). In addition, in Table 2, both fixed and random effects results are shared.

Test of Homogeneity

In the homogeneity test, Hedges's Q homogeneity test was used. Significant Q values show that there is a significant difference between steps of the independent variable. In other words, it shows that observed studies come from populations more than one, its variety and heterogeneity (Hedges & Vevea, 1998). In the meantime, I^2 statistics were used in interpreting homogeneity/heterogeneity. I^2 statistics range between 0 and 100%. I^2 statistics are interpreted as 25%, 50% and 75%, meaning as low, average and high, respectively (Higgins et al., 2003).

Test for Moderator Effects

The subject area-matter test content could affect scores obtained from computer-based and paper-and-pencil forms. Thus, in this study, it was tested whether mean effect sizes change or not depending on sub-groups for some variables. In terms of the equivalence of the PP and CBT forms, it was investigated whether mean effect sizes differ or not according to variables like type of computerized, education level, and subject matter. In this direction, ANOVA and Q values were used.

Publication Bias

To decrease publication bias, all studies, dissertations, articles, reports related to the subject and containing Turkey sample in accordance with the purpose of the study and criteria of choosing were tried to be achieved. In addition, in examining the effects of publication bias, several methods were used. In line with this, Funnel plot graphics, Begg and Mazumdar's rank correlation test, Egger's regression method, Duval and Tweedie's trim and fill method, Rosenthal's fail-safe N tests results were examined. In addition, the normality of the effect sizes of the studies was examined and it was determined that they showed a normal distribution. In the meta-analysis study, the Comprehensive Meta-Analysis, Version 3 (CMA; Borenstein, Hedges, Higgins, & Rothstein, 2014) program was used.

Results

Characteristics of the Included Studies

Descriptive statistics related to the studies included in the meta-analysis are given in Table 1.

Table 1

Characteristics of the Included Studies

Type of Research	N	%	*Education Level*	N	%
Master's Thesis	14	37.84	Secondary School	10	27.03
Doctoral Dissertation	13	35.13	High School	9	24.32
Article	10	27.03	University	13	35.13
Total	37	100	Others	5	13.52
Publication Year	N	%	Total	37	100
1993-1999	2	5.41	*Subject Matter*	N	%
2000-2006	3	8.11	Quantitative	4	10.81
2007-2013	20	54.05	Verbal	5	13.51
2014-2020	12	32.43	Science	6	16.22
Total	37	100	English	3	8.11
Type of Computerized	N	%	Computer Achievement	3	8.11
			Psychological and Diagnostic	13	35.13
CAT	14	37.84			
CBT-P	23	62.16	Others	3	8.11
Total	37	100	Total	37	100

When characteristics of the studies included in Table 1 were investigated, it was seen that 14 of the studies (37.84%) weremaster's thesis, 13 of them (35.13%) were doctoral dissertations and 10 of them (27.03%) were articles. 54.05% of the studies were conducted between 2007 and 2013, 32.43% of them were conducted between 2014 and 2020. Before 2000, there were scarcely few studies (5.41%) comparing computer-based and paper-and-pencil examinations. However, in recent years, studies focusing on this subject increased. When CBT classification was examined, it could be seen that 23 studies (62.16%) were computerized implementation of paper-and-pencil tests (CBT-P) and 14 studies (37.83%) were CAT.

When the educational level of students in meta-analysis was examined, it was found that 10 studies (27.03%) were conducted with secondary school students, nine studies (24.32%) were conducted with high school students, 13 studies (35.13%) were conducted with university students and five studies (13.52%) were conducted with students who were ill or attending driving license examination and over 18 years old.

In studies included in the meta-analysis, PP and CBT comparisons were made in different subject matters. For PP and CBT comparisons, 10.81% of them were made in quantitative and 13.51% of them in verbal, 16.22% in science and 8.11% in English courses. 8.11% of the studies were conducted on subjects like computer class achievement and achievement in software. 35.13% of the studies were conducted on matters to detect "psychological and diagnostic" students' self-competence, anxiety, attitude, line orientation test, diagnosing/detecting musculoskeletal problems. In addition, 8.11% of the studies were conducted on subjects like spatial ability, instructional design, driving license exam, measurement and evaluation and success of evaluation described as others.

Homogeneity Test

For the homogeneity test, Q statistics and p-value were examined. Homogeneity test results are given in Table 2.

Table 2

Summary of Effect Sizes and Homogeneity Test

			95% Interval				Homogeneity			
Model	N	Effect Size	Lower Limit	Upper Limit	Z	*p*	Q	*df*	*p*	I^2
Fixed	37	0.037	0.032	0.041	15,4	0.00.0	577.79	36	0.0	93.8
Random	37	0.042	0.005	0.079	2.215	27				

In heterogeneity results given in Table 2, $Q_{(36)}$=577.79, $p < .001$, indicating that there was heterogeneity. The I^2 is 93.8% of the observed variance in effects. The I^2 is 93.8%, which demonstrated a high amount of heterogeneity. This tells us that the true effect size probably varies across studies, which means that the data are not consistent with the assumptions of the fixed-effect model (Borenstein, Hedges, Higgins & Rothstein, 2015).

In addition, in Table 2, effect sizes were given according to the fixed-effect model and random effect model. According to the fixed-effect model, the average effect size value is 0.037 (ranged from 0.032-0.041) for a 95% confidence interval, and according to the random effect model, it is 0.042 (ranged from 0.005-0.079). The random effect had a Z-value 2.215 ($p < .05$). Thus, we can reject the null hypotheses that the true mean difference is 0.0.

Publication Bias

For studies used in this research, publication bias was tested with different methods. First of all, the funnel plot was examined and shown in Figure 1. According to Figure 1, that the funnel plot showed a relatively symmetrical distribution shows that there was no publication bias.

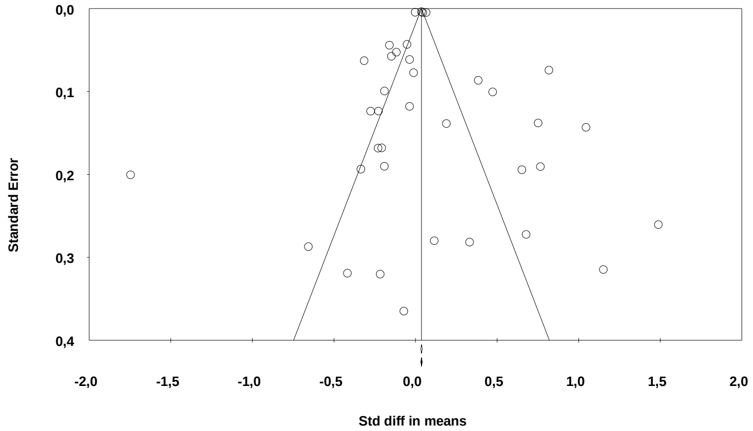


Figure 1. *Funnel Plot*

Interpretation of funnel plot is subjective (Borenstein et al., 2009). Thus, the process must be supported by other evidence. In this direction, first of all, Begg and Mazumdar's rank correlation test results were examined. Begg's test results show that there is no publication bias (tau b 0.10210, *p* > .05). Thereafter, Egger's regression test results were examined. Egger's regression test results show that there is no publication bias (Intercept is 0.20217 (95% CI=-1.30296-1.70730, *p* > .05). Rosenthal's fail-safe N was 534. This means that we would need to locate and include 534 'null' studies in order for the combined 2-tailed p-value to exceed 0.050. The value of 534 was much larger than the value of 195 (5k+10 formula). Thus, there was no publication bias in the findings. Duval and Tweedie's trim and fill test method was used and results are given in Table 3.

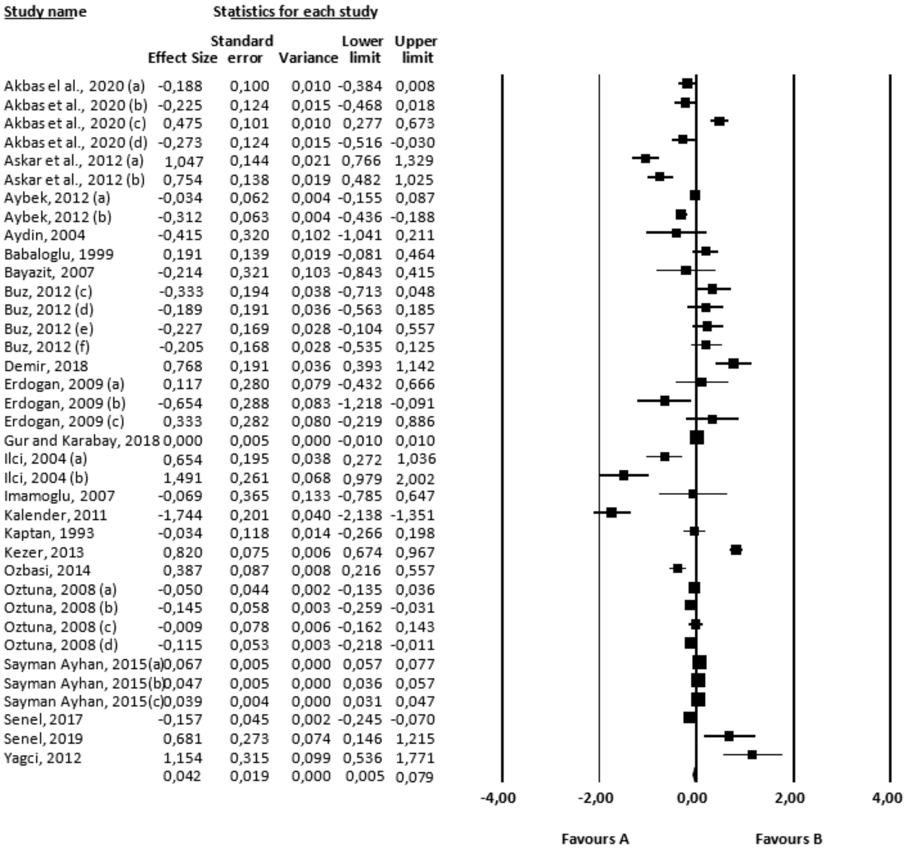
Table 3

The Result of Duval and Tweedie's Trim and Fill Test

	Studies Trimmed (right and left)	Point Estimate	Lower Limit	Upper Limit	Q
Obs. Values		0.04170	0.00481	0.07859	577.78884
Adj. Values	0	0.04170	0.00481	0.07859	577.78884

Just as given in Table 3, in this study, there was no difference between observed effect size value (0.04170) and adjusted effect size (0.04170), which wascreated to correct the effect emerging from publication bias, according to random effect model. In addition, according to random effect model, the trimmed value seems to be zero. As a result, it can be said that the average effect size is not a result of publication bias.

A total of 37 effect sizes from 21 studies were estimated for the current study. The forest plot graphic, which shows the meta-analysis results of the research, is given in Figure 2.



Study name	Statistics for each study				
	Effect Size	Standard error	Variance	Lower limit	Upper limit
Akbas el al., 2020 (a)	-0,188	0,100	0,010	-0,384	0,008
Akbas et al., 2020 (b)	-0,225	0,124	0,015	-0,468	0,018
Akbas et al., 2020 (c)	0,475	0,101	0,010	0,277	0,673
Akbas et al., 2020 (d)	-0,273	0,124	0,015	-0,516	-0,030
Askar et al., 2012 (a)	1,047	0,144	0,021	0,766	1,329
Askar et al., 2012 (b)	0,754	0,138	0,019	0,482	1,025
Aybek, 2012 (a)	-0,034	0,062	0,004	-0,155	0,087
Aybek, 2012 (b)	-0,312	0,063	0,004	-0,436	-0,188
Aydin, 2004	-0,415	0,320	0,102	-1,041	0,211
Babaloglu, 1999	0,191	0,139	0,019	-0,081	0,464
Bayazit, 2007	-0,214	0,321	0,103	-0,843	0,415
Buz, 2012 (c)	-0,333	0,194	0,038	-0,713	0,048
Buz, 2012 (d)	-0,189	0,191	0,036	-0,563	0,185
Buz, 2012 (e)	-0,227	0,169	0,028	-0,104	0,557
Buz, 2012 (f)	-0,205	0,168	0,028	-0,535	0,125
Demir, 2018	0,768	0,191	0,036	0,393	1,142
Erdogan, 2009 (a)	0,117	0,280	0,079	-0,432	0,666
Erdogan, 2009 (b)	-0,654	0,288	0,083	-1,218	-0,091
Erdogan, 2009 (c)	0,333	0,282	0,080	-0,219	0,886
Gur and Karabay, 2018	0,000	0,005	0,000	-0,010	0,010
Ilci, 2004 (a)	0,654	0,195	0,038	0,272	1,036
Ilci, 2004 (b)	1,491	0,261	0,068	0,979	2,002
Imamoglu, 2007	-0,069	0,365	0,133	-0,785	0,647
Kalender, 2011	-1,744	0,201	0,040	-2,138	-1,351
Kaptan, 1993	-0,034	0,118	0,014	-0,266	0,198
Kezer, 2013	0,820	0,075	0,006	0,674	0,967
Ozbasi, 2014	0,387	0,087	0,008	0,216	0,557
Oztuna, 2008 (a)	-0,050	0,044	0,002	-0,135	0,036
Oztuna, 2008 (b)	-0,145	0,058	0,003	-0,259	-0,031
Oztuna, 2008 (c)	-0,009	0,078	0,006	-0,162	0,143
Oztuna, 2008 (d)	-0,115	0,053	0,003	-0,218	-0,011
Sayman Ayhan, 2015(a)	0,067	0,005	0,000	0,057	0,077
Sayman Ayhan, 2015(b)	0,047	0,005	0,000	0,036	0,057
Sayman Ayhan, 2015(c)	0,039	0,004	0,000	0,031	0,047
Senel, 2017	-0,157	0,045	0,002	-0,245	-0,070
Senel, 2019	0,681	0,273	0,074	0,146	1,215
Yagci, 2012	1,154	0,315	0,099	0,536	1,771
	0,042	0,019	0,000	0,005	0,079

Figure 2. *Forest Plot*

As could be seen in Figure 2, according to the random effect model, with 0.019 standard error and 95% confidence interval, the effect size (mean ES) value was 0.042 (ranged from 0.005-0.079). In addition, the Hedges' g effect size is found as 0.042. According to Cohen's classification, this value revealed that the test method (PP or

CBT) had a negligible effect. Effect size close to zero indicated the equivalence of the standardized means of the PP and CBT forms. The positive effect size indicated that the CBT score on average was slightly higher than PP (ES =CBT-PP). When effect sizes of the studies were examined, it could be seen that out of 37 studies, 17 had a positive effect and 20 had a negative effect.

Moderator Analysis

Concerning the equivalence of the PP and CBT forms, findings of whether effect sizes varied or not according to some moderator variables (type of computerized, education level, subject matter) are given in Table 4.

Table 4

Results of Moderator Analysis

Moderator Variable	N	ES	Lower Limit	Upper Limit	Q	*df*	*p*
Types of Computerized					0.712	1	0.399
CAT	14	0.031	-0.013	0.076			
CBT-P	23	0.066	-0.001	0.133			
Education Level					65.780	3	0.00
Secondary School	10	-0.173	-0.255	-0.090			
High School	9	0.125	0.059	0.191			
University	13	0.244	0,166	0.322			
Others	5	-0.057	-0.126	0.012			
Subject Matter					86.173	6	0.00
Quantitative	4	0.068	-0.054	0.189			
Verbal	5	-0.167	-0.268	-0.067			
Science	6	-0,009	-0.074	0.056			
English	3	0.676	0.501	0.851			
Computer Achievement	3	0.424	0.239	0.610			
Psychological and Diagnostic	13	0.039	-0.024	0.102			
Others	3	0.022	-0.091	0.136			

As can be seen in Table 4, the difference between the effect sizes of the type of computerized method was not statistically significant (Q= 0.712, *p* > .05). There was no evidence that the effect varies by the type of computerized method. In other words, different types of computerized methods (CAT or CBT-P) yielded similar results. The mean ES values were 0.031 and 0.066 for CAT and CBT-P, respectively. CAT had a positive mean ES and also CBT-P had a positive mean ES. This means that students had slightly higher scores for CAT and CBT-P than PP. However, both of these ESs indicated statistically significant equivalence between both modes of CBT and PP.

In the moderator analysis performed, the mean ES values were -0.173, 0.125, 0.244, and -0.057 for secondary school, high school, university, and others, respectively. The difference between the effect sizes of the education levels was statistically significant (Q= 65.780, p < .01). There was evidence that the impact of PP and CBT comparisons varied by education level. According to these findings, scores obtained as a result of PP or CBT implementations from different educational levels varied statistically significant . In this difference, the effect size of university students was 0.244. Therefore, the PP and CBT differences appeared more in this group than the others.

The difference between the effect sizes of the subject matters was statistically significant (Q=86.173, p < .01). There was evidence showing that the effects of PP and CBT comparisons varied by the subject matter. The effect size for the CBT-PP difference was medium in English and small in computer measurements.

Discussion, Conclusion and Recommendations

The present study aims to investigate the effects of the difference between PP and CBT implementations in Turkey. In this direction, between 1993 and 2020, a meta-analysis was conducted with the findings of 37 studies with samples from Turkey. As a result of meta-analysis conducted, common effect size was 0.042. The mean effect size found was negligible. In this direction, it was found that the difference in test implementation methods (CBT-PP) was negligible. In this case, it can be said that students' performance on CBT is not significantly better than their performance on PP. This result which includes comparison studies conducted between 1993 and 2020 in Turkey is parallel with findings obtained in other meta-analysis studies between varied years, culture and subjects (Aybek et al., 2014; Bergstrom, 1992; Finger & Ones, 1999; Kingston, 2009; Kim, 1999; Mead & Drasgow, 1993; Wang et al., 2007; Wang et al., 2008). In some individual studies investigating PP and CBT scores/abilities on varying subjects, it is frequently discussed that there is a positive and strong relationship between PP and CBT results (Aybek, 2016; Aytug-Kosan, 2013; Bulut & Kan, 2012; Iseri, 2002; Kaskati, 2011; Kim & Huynh, 2008; Kingsbury, 2002; Sahin, 2017; Simsek, 2017). Based on the research results and literature findings, it can be said that CBT and PP applications give similar results. Accordingly, these results are expected to shed light on measurement practices in distance education in the pandemic period (COVID-19) on a national basis.

Also, in this study, sub-group examinations were conducted according to some variables. It was identified that the type of computerized method was not moderators in the differences in mean scores between the PP and CBT. Moreover, we can say that the effect size does not differ by the types of computerized methods. There is no evidence that any of the type of computerized methods (CAT or CBT-P) is more effective than the other. Thus, we can say that the types of CBT versions of CAT and CBT-P are equivalent to PP in this research. Contrary to this result obtained from the study, Kim (1999), in comparison of CAT and CBT-P test results, found a statistical difference. Schaeffer et al. (1995) found that in their study comparing scores of computer-based and CAT versions of the GRE test, verbal and quantitative results were comparable. However, in the same study, it was found that analytical CAT and computer-based scores were not comparable in the analytical test. The differences can be thought to decrease with the development of software. Studies taking part in the study of Kim (1999) goes to the year 1996 and before and studies taking part in Schaeffer's (1995) study goes back to 1995 and before. This study usually goes to 1996 and after. Variables like the development of computer systems and software over time may have minimized differences and problems based on the CAT and CBT-P.

There is an evidence that the effect varied by education level (secondary school, high school, university and over 18 years old). Education level leads to differences in student scores between computer-based and paper-and-pencil tests. Kim (1999), in the meta-analysis study conducted using the Gleser-Olkin method, found that PP-CAT and PP-CBT comparisons did not differ according to high school or college. Kim (1999), in the typical meta-analytical study, found that there was no difference between CAT and PP concerning educational levels, but there was a difference between computer-based-PP according to educational levels. Kingston (2009), in comparison to PP and CBT difference, found that there is no significant difference in comparison of confidence intervals of education levels (elementary, middle, high), which were used as moderator effect. Wang et al. (2007) found that educational level did not lead to differences in students' mathematics mean scores between computer-based and paper-and-pencil modes. Wang et a.l (2008) found that education level did not affect the differences in reading scores between test modes. The difference in research results can be explained by the lack of computer ownership and/or the level of computer usage and familiarity in the Turkey sample. In addition, the difference found in education levels in this study is statistically significant but not important because effect size values according to education level are very close to each other.

Regarding research findings, the effect of PP and CBT implementation designs on student score/ability showed a meaningful difference according to some determined subjects. According to Kim's (1999) study, CAT versions for mathematics and other cognitive measurements were equivalent to PP versions, while CAT versions for English tests and other subjects' tests were not. Kingston (2009), in the meta-analysis study conducted on PP and CBT comparison, found subject matter comparisons statistically significant however, stated that a great part of variability is not explained by the subject matter. In this study, as in the literature, some subject matters lead to differences in student scores between PP and CBT. Especially, this difference is more in English and Computer achievement tests. At present, especially to measure language skills (e.g., English), TOEFL, IELTS, CBT implementations are quite common. Also, in measuring skills like writing, verbal reasoning, quantitative reasoning, CBT implementations can be seen, just like GRE. Accordingly, concerning subjects investigated in this research, it can be said that CBT implementations at a national level can be used in scales measuring quantitative, science, and used for psychological and diagnostic purposes.

This study is limited to three moderator effects that are thought to affect the means effect size. Different moderators like characteristics of computer-based test systems, testing environment, cultures, countries and sex may be examined. Also, in another study, to reveal cultural differences, an inter-cultural comparison may be held. The effects related to PP and CBT differences in different countries/cultures can be compared. PP and CBT results of students in different school environments concerning technical characteristics can be compared.

Results obtained from the research provided more evidence to support the comparability of PP and CBT. Results obtained show CBT can be an acceptable alternative to traditional pencil and paper tests. Also, in the study effect size of the CBT-PP difference was found negligible in high school and primary education. In addition, the mean effect size of the CBT-PP difference was negligible also in quantitative, verbal, science and psychological measurement. Accordingly, it can be recommended that CBT can be used in quantitative, verbal, science and psychological measurement instead of PP in high school and primary education in Turkey. In this way, results obtained are expected to lead national educational policies and guide studies and national measurement implementations in the future.

This study is limited to studies with Turkish sample, conducted between 1993 and 2020, implemented on a computer-based/computer-adaptive and paper-and-pencil format, and could be accessed by the researcher. As a result, it can be said that the results of this study will be utilized for Turkey. Also, although there are no statistical results in findings showing a publication bias, in accordance with the purpose of this study, studies only reporting relationships between PP and CBT success/ability were excluded (n=3). Additionally, access to unpublished documents was limited to dissertations.

References

(The studies with an asterisk (*) indicate that they are used in this meta-analysis research.)

*Akbas, U., Aydogdu, S., & Buyukozturk, S. (2020). Investigation of psychometric traits of metric scale and likert type scale applied in different conditions. *Hacettepe University Journal of Education*, 35(1), 222-242. doi: 10.16986/HUJE.2019050088.

*Askar, P., Altun, A., & Cangoz, B. (2012). A comparison of paper-and-pencil and computerized forms of line orientaion and enhanced cued recall tests. *Psychological Reports*, 110(2), 383-396.

*Aybek, E. C. (2012). *A comparison of psychometric properties of a general ability test which administered in paper-pencil and computer based form* [Master's thesis, Ankara University]. Council of Higher Education Thesis Center.

Aybek, E., C., Sahin, D., B., Eris, H. M., Simsek, A. S., & Kose, M. (2014). Meta-analysis of comparative studies of student achievement on paper-pencil and computer-based test. *Asian Journal of Instruction*, 2(2), 18-26.

Aybek, E. C. (2016). *An investigation of applicability of the self-assessment inventory as a computerized adaptive test* [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center.

*Aydin, S. (2004). *The efficiency of computers on testing writing skills* [Doctoral dissertation, Atatürk University]. Council of Higher Education Thesis Center.

Aytug-Kosan, M.A. (2013). *Developing an item bank for progress test and application of a computerized adaptive testing by simulation in medical education* [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center.

*Babaloglu, N. (1999). *Comparing the psychometric properties that obtained from the multiple choice test applied with paper-pencil test and computer based application* [Unpublished master's thesis]. Hacettepe University.

*Bayazit, A. (2007). *Testing time and performance differences between online and paper-pencil tests* [Master's thesis, Hacettepe University]. Council of Higher Education Thesis Center.

Berben, L., Sereika S. M., & Engberg, S. (2012). Effect size estimation: methods and examples. *International Journal of Nursing Studies*, 49(8), 1039-47.

Bergstrom, B. (1992). Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis. *Paper presented at the annual meeting of the American Educational Research Association*, San Francisco, CA. https://eric.ed.gov/?id=ED377228

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (2009). *Introduction to meta-analysis*. United Kingdom: John Wiley & Sosns, Ltd., Publication.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2014). *Comprehensive meta-analysis version 3.* Englewood, NJ: Biostat.

Borensten, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2015). Regression in Meta-Analysis. Available at https://www.meta-analysis.com/downloads/MRManual.pdf

Bulut, O., & Kan, A. (2012) Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eurasian Journal of Educational Research*, 49, 61–80.

*Buz, G. (2012). *Success analysis web based survey assessment and evaluation* [Master's thesis, Beykent University]. Council of Higher Education Thesis Center.

Coe, R. (2002). It's the effect size, stupid. What effect size is and why it is important. *In Proceedings of the British Educational Research Association*, 12-14 September 2002, England. Available at https://www.leeds.ac.uk/educol/documents/00002182.htm

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cikrikci-Demirtasli, N. (1995) Psikometride yeni ufuklar: Bilgisayar ortamında bireye uyarlanmış test. *Türk Psikoloji Bülteni*, 5 (13), 31-36.

*Demir, S. (2018). *Investigation of different item selection methods in terms of stopping rules in polytomous computerized adaptive testing* [Doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis Center.

*Erdogan, Y. (2009). Paper-based and computer-based concept mappings: The effects on computer achievement, computer anxiety and computer attitude. *British Journal of Educational Technology*, 40(5), 821–836.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, 59 (4), 395-430.

Finger, M. S., & Ones, D. S. (1999). Psychometric equivalence of the computer and booklet forms of the MMPI: A meta-analysis. *Psychological Assessment*, 11(1), 58-66.

Goldberg, A., Russell, M., & Cook, A. (2003). The Effect of Computers on Student Writing: A Meta-analysis of Studies from 1992 to 2002. *The Journal of Technology, Learning and Assessment*, *2*(1). Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1661

*Gur, R., & Karabay, E. (2018). Simulative computerized adaptive motor vehicle driving license exam test administration. *Journal of Theoretical Educational Science*, 11(2), 201-228.

Hambleton, R. K., Zaal, J. N., & Pieters, J. P. M. (1991). Computerized adaptive testing: Theory, applications, and standards. In R. K. Hambleton & J. N. Zaal (Eds.), *Evaluation in education and human services series. Advances in educational and psychological testing: Theory and applications* (p.341–366). Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/978-94-009-2195-5_12

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods,* 3, 486-504.

*Ilci, B. (2004). *To compare the psychometric properties that obtained from the measurement of numerical abilities and verbal abilities with multiple choice tests through paper-pencil test and online application on the computer* [Unpublished master's thesis]. Hacettepe University.

*Imamoglu, C. (2007). *Computer based versus paper based assessment in English language teaching* [Master's thesis, Marmara University]. Council of Higher Education Thesis Center.

Iseri, A. I. (2002). *Assessment of students' mathematics achievement through computer adaptive testing procedures* [Unpublished doctoral dissertation]. Middle East Technical University.

*Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability* [Doctoral dissertation, Middle East Technical University]. Council of Higher Education Thesis Center.

*Kaptan, F. (1993). *Comparison of adaptive test application and traditional paper-pencil test application in ability estimation* [Doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis Center.

Kaskati, O. T. (2011). *Development of computer adaptive testing method using with rasch models for assessment of disability in rheumatoıd arthrıtıs patients* [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center.

*Kezer, F. (2013). *Comparıson of the computerized adaptive testing strategies* [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center.

Kim, J. (1999). Meta-analysis of equivalence of computerized and P&P tests on ability measures. *In Annual Meeting of the Mid-Western Educational Research Association*, Chicago, IL. https://files.eric.ed.gov/fulltext/ED449182.pdf

Kim, D. H., & Huynh, H. (2008). Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course English test. *Educational and Psychological Measurement*, 68(4), 554-570.

Kingsbury, G. G. (2002). An empirical comparison of achievement level estimates from adaptive tests and paper-and-pencil tests. *In Annual Meeting of the American Educational Research Association*, New Orleans, LA. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.617.7215&rep=rep1&type=pdf

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.

Marfo, P., & Okyere, G. A. (2019). The accuracy of effect-size estimates under normals and contaminated normals in meta-analysis. *Heliyon*, 5(6), 1-9. https://doi.org/10.1016/j.heliyon.2019.e01838

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods*, 11(4), 386-401.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 9, 287-304.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*(1), 105–125. https://doi.org/10.1037/1082-989X.7.1.105

*Ozbasi, D. (2014). *Research on applicability of computer literacy test as computerized adaptive testing* [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center.

*Oztuna, D. (2008). *An application of computerized adaptive testing in the evaluation of disability in musculoskeletal disorders* [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center.

*Sayman Ayhan, A. (2015). *Comparability of scores from cat and paper and pencil implementations of student selection examination to higher education* [Master's thesis, İhsan Doğramacı Bilkent University]. Council of Higher Education Thesis Center.

Schaeffer, G. A., Steffen, M., Golub-Smith, M.L., Mills, C.N. & Durso, R. (1995). *The introduction and comparability of the computer adaptive GRE general test* (GRE Board Professional Report No. 88-08aP). Princeton, NJ: Educational Testing Service.

Sun, R. W. & Cheung, S. F. (2020). The influence of nonnormality from primary studies on the standardized mean difference in meta-analysis. *Behavior Research Methods*, https://doi.org/10.3758/s13428-019-01334-x

Sahin, M. D. (2017). *Examining the results of multidimensional computerized adaptive testing applications in real and generated data sets* [Doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis Center.

*Senel, S. (2017). *Investigation of the compatibility of computerized adaptive testing on students with visually impaired* [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center.

*Senel, H. C. (2019). *Comparing the effect of tablet, desktop, paper-pencil based drill practices on spatial skills of vocational high school students* [Doctoral dissertation, METU]. Council of Higher Education Thesis Center.

Simsek, A. S. (2017). *Adaptation of skills confidence vocational interest inventory and development of computerized adaptive test application* [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center.

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

Wang, S., Jiao, H., Young, M. J., Brooks, T. & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments. *Educational and Psychological Measurement*, 68(1), 5-24.

Wolf, F. M. (1986). *Meta-Analysis: Quantitative methods for research synthesis*. Sage Publication: London.

*Yagci, M. (2012). *Designing a new online examination model and a comparison with paper-based test* [Doctoral dissertation, Sakarya University]. Council of Higher Education Thesis Center.

Kağıt-Kalem Formuda ve Bilgisayar Ortamında Uygulanan Testlerin Karşılaştırılması: Türkiye Örnekleminde Bir Meta-Analiz Çalışması

Atıf:

Nalbantoğlu Yılmaz, F. (2021). Comparing paper-pencil and computer-based tests: A meta-analysis study in the sample of Turkey. *Eurasian Journal of Educational Research*, 93, 279-300, DOI: 10.14689/ejer.2021.93.13

Özet

Problem Durumu: Bilgisayar teknolojisindeki ilerlemeler yaşantımızın birçok noktasında bilgi teknolojilerini kullanmayı artırmakta ve hatta gerekli kılmaktadır. Bilgisayarlar yaygın olarak günlük işlerde olduğu gibi eğitimin her noktasında da sıklıkla kullanılmaktadır. Eğitimde bilgisayar teknolojileri öğrencileri tanımada, öğrencilere ilişkin verilerin saklanmasında olduğu gibi öğretim ve ölçme ve değerlendirme süreçlerinde de yaygın olarak kullanılmaktadır. Bilgisayar teknolojilerinin yaşantımıza bu denli girmesi bilgisayar ortamindaki test uygulamalarının artmasına olanak sağlamaktadır. Bilgisayar ortamında uygulanan testler (CBT), kağıt-kalem testlerinin bilgisayarda ortamında uygulanması (CBT-P) veya bilgisayar ortamında bireye uyarlanmış (CAT) testler olarak iki biçimde uygulanabilmektedir.

Dünyada, bilgi teknolojilerindeki ve bilgisayar kullanımındaki gelişimle birlikte bilgisayar ortamında ve/veya bilgisayar ortamında bireye uyarlanmış test uygulamalarının giderek arttığı görülmektedir. Fakat bilgisayar ortamında test uygulamaları, özellikle de bilgisayar ortamında bireye uyarlanmış test uygulamalarının Türkiye açısından hala sınırlı kaldığı bir gerçektir. Bilgisayar teknolojilerindeki ilerlemeler ve bilgisayar ortamındaki test uygulanmaları, bilgisayar ortamında bireye uyarlanmış test uygulamalarının avantajlı noktaları düşünüldüğünde, kağıt-kalem ve bilgisayar ortamında uygulanan sınavların psikometrik özelliklerinin, öğrencilerin başarılarının karşılaştırılması gerekliliği ise kaçınılmazdır. Örneğin; her iki uygulamadan elde edilen puanlar farklılık göstermekte midir? Her iki uygulamadan elde edilen yetenek kestirimleri benzer midir?

Günümüzde, kağıt-kalem testleri ile bilgisayar ortamında bireye uygulanan testlerin karşılaştırıldığı çalışmalar giderek artmaktadır. Yapılan çalışmalardan elde edilen bilgi birikimlerini birleştirerek yorumlamak ise günümüz test uygulayıcılarına, psikometristlere önemli bilgiler sunacaktır.

Literatürde bilgisayar ortamında ve kağıt kalem formatında yapılan test uygulamalarını karşılaştırıldığı bir çok çalışma mevcuttur. Bu çalışma sonuçlarının irdelenmesi bilgisayar ortamında bireye uyarlanmış test ve kağıt-kalem test uygulamalarına katkı getireceği düşünülmektedir. Bu doğrultuda araştırmada, Türkiye'deki bu konuda yapılmış çalışmalar bir takım istatistiksel veriler yardımıyla birleştirilerek sistematik şekilde özetlenmek istenmektedir. Bu amaçla araştırmada

aynı konudaki farklı çalışma sonuçlarını niceliksel olarak sentezleme çalışması olan meta-analiz kullanılmıştır.

Eğitimde teknoloji kullanımı, bilgisayar aşinalığı vb. etkiler ülke hatta bölgeler bazında dikkate alındığında PP ve CBT formunda uygulanan sınavlardaki başarının/yeteneğin karşılaştırıldığı bu meta analiz çalışmasının Türkiye örneklemiyle gerçekleştirilen çalışmaları kapsaması önemli görülmektedir. Çünkü bilgisayar ortamındaki test uygulamasında bireylerin bilgisayar aşinalığı, bilgisayar yeterliliği vb. durumlar ülke hatta bölge bazında farklılık gösterebilir.

Araştırmanın Amacı: Araştırmanın amacı, 1993-2020 yılları arasındaki Türkiye örneklemini içeren çalışmalardan hareketle PP ve CBT arasındaki ortalama farkların etkisini meta analiz kullanarak belirlemektir. Bu sayede araştırma sonuçlarının, Türkiye'de sıklıkla uygulanan, bireyler hakkında önemli kararların alındığı, geniş ölçekli testlerin uygulanma yolu/metodu ile ilgili ölçme uygulamamalarına ve alana katkı sağlaması beklenmektedir. Ayrıca dünyada yaşanan pandemi sırasında uzaktan eğitimde çevrimiçi sınavların kağıt-kalem testlerine göre güvenirliği, benzerliği tartışma konusu olmuştur. Bu aşamada, yapılan karşılaştırmaları içeren meta analiz çalışmasının önemli bilgiler vereceği beklenilmektedir.

Araştırmanın Yöntemi: Araştırma, Türkiye örnekleminde 1993-2020 yılları arasındaki bilgisayar ortamında ve kağıt-kalem formunda yapılan ölçme uygulamalarının karşılaştırıldığı çalışmaları içeren bir meta-analiz çalışmasıdır. Yapılan detaylı literatür taramasından elde edilen çalışmalar tek tek incelenmiş ve seçme kriterleri de dikkate alınarak araştırma kapsamına giren çalışmalar belirlenmiştir. 37 bulgu meta-analize dahil edilmiştir. Çalışmada bazı değişkenler için ortalama etki büyüklüklerinin alt gruplara göre değişip değişmediği de test edilmiştir. Bu doğrultuda ANOVA ve Q değeri kullanılmıştır. Yayın yanlılığını azaltmak için, konu ile ilgili tezler, makaleler, raporlar olmak üzere araştırma amacı ve seçim kriterleri doğrultusunda Türkiye örneklemini içeren tüm çalışmalara ulaşılmaya çalışılmıştır. Bununla birlikte yayın yanlılığı etkilerinin incelenmesinde Funnel plot grafiği, Begg's testi, Egger's regression methodu, Duval and Tweedie's trim and fill methodu, Rosenthal's fail-safe N testi kullanılmıştır.

Araştırmanın Bulguları: Araştırmada rastgele etki modeline göre hesaplanan ortalama etki büyüklüğü değeri 0.042 olarak belirlenmiştir. Yapılan yanlılık incelemeleri sonucunda elde edilen etki büyüklüğünün yayın yanlılığının sonucu olmadığı belirlenmiştir. PP ve CBT farklılığına ilişkin elde edilen ortalama etki büyüklüğü bilgisayar yöntemine göre istatistiksel olarak anlamlı farklılık göstermemektedir.

Araştırmanın Sonuçları ve Öneriler: Araştırmadan elde edilen sonuçlar PP ve CBT'nin karşılaştırılabilirliğini destekleyen daha fazla kanıt sağlamıştır. Elde edilen sonuçlar CBT'nin geleneksel kalem ve kağıt testlerine kabul edilebilir bir alternatif olabileceğini göstermektedir. Elde edilen sonuçların ulusal eğitim politikalarına yön vermesi, gelecekteki çalışmalara ve ulusal ölçme uygulamalarına rehberlik etmesi beklenmektedir.

PP ve CBT uygulamalarının öğrenci başarısı/yeteneği üzerindeki etkisi bilgisayarlı yöntemin türüne göre anlamlı bir farklılık göstermemiştir. Bilgisayar sistemlerinin ve yazılımın zamanla gelişmesi, insanların bilgisayar okuryazarlığının gelişmesi, bilgisayar kullanım yaşının düşmesi gibi etkiler farklılıkları ve sorunları en aza indirmiş olabilir.

Bu çalışma, ortalamaların etki boyutunu etkilediği düşünülen üç değişken etkisi ile sınırlıdır. Bilgisayar tabanlı test sistemlerinin özellikleri, test ortamı, kültürler, ülkeler, cinsiyet vb. gibi farklı etkilerde başka bir çalışmada incelenebilir. Ayrıca, kültürel farklılıkları ortaya çıkarmak için, kültürler arası bir karşılaştırma yapılabilir. Farklı ülkelerdeki / kültürlerdeki PP ve CBT farklılıklarına ilişkin etkiler karşılaştırılabilir. Farklı okul ortamlarındaki öğrencilerin PP ve CBT sonuçları teknik özellikleri açısından karşılaştırılabilir. Bu çalışma, 1993 ve 2020 yılları arasında, bilgisayar tabanlı/bilgisayar uyarlamalı ve kâğıt-kalem formatında uygulanan ve araştırmacı tarafından erişilebilir olan Türkiye örneklemli çalışmalarla sınırlıdır. Araştırma sonuçlarının Türkiye için kullanılacağı söylenebilir.

Anahtar Sözcükler: Kağıt-kalem testleri, bilgisayar ortamında uygulanan test, etki büyüklüğü, meta-analiz.