Check for updates

Assessment
1–12
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10731911221100995
journals.sagepub.com/home/asm
SAGE

The Effect of Computerized Adaptive Testing on Motivation and Anxiety: A Systematic Review and Meta-Analysis

Hanif Akhtar[1,2], Silfiasari[2], Boglarka Vekety[1], and Kristof Kovacs[1]

Abstract

Although many studies have been carried out on the psychometric aspects of computerized adaptive testing (CAT), its psychological aspects are less researched. Early studies claimed that CAT can be more motivating and induce less anxiety than traditional fixed-item tests (FIT). The purpose of this systematic review and meta-analysis was to gain a comprehensive understanding of the effects of CAT on motivation and anxiety in comparison to traditional fixed-item testing. Seven databases were examined. Articles were eligible if they employed an empirical study containing a direct comparison between CAT and FIT. Meta-analytical results showed no overall effect of test type on anxiety and motivation when comparing CAT with FIT ($k = 11$, $g+ = 0.06$, $p = .28$). However, easier CAT had positive effect compared with FIT ($k = 2$, $g+ = .22$, $p < .001$). Certain modifications in CAT administration can provide positive psychological effects for test-takers.

Keywords

computerized adaptive testing, fixed-item testing, motivation, anxiety, meta-analysis

Computerized adaptive testing (CAT), a method with increasing popularity, has two pillars: computer technology and item response theory (IRT). The basic concept behind CAT is that test items are selected by an algorithm to match each test-taker's ability level (Wainer, 2000). CAT is a branch and the further development of computer-based testing (CBT). Although in traditional CBT every test-taker is presented with the same set of items, the selection of items in CAT is tailored to the individual test-taker's performance. The test-taker's ability level is iteratively estimated during the testing process and items are presented based on the current ability estimate, which depends on the examinees' previous answers. Hence, different test-takers are presented with different items.

If the test-taker correctly answers the first item on a CAT, a more difficult item follows. If the test-taker provides an incorrect answer, the next item is easier. As a result, the difficulty of each item administered after the first one is determined by the result of all previously administered items. Items that are too easy or too difficult for test-takers contribute very little information about their ability, therefore test-takers are usually only receive items that have a success probability of nearly 50% (Green et al., 1984; van der Linden & Pashley, 2000). Testing continues until a stopping criterion is met (e.g., the test exceeds the predetermined number of items, or until the standard error falls below a predetermined threshold). The standard error of measurement (SEM) decreases after each item, as increasing information is provided of the examinee's ability. This mechanism makes it possible to decrease the number of items administered without sacrificing precision (Lunz et al., 1994; Wainer & Eignor, 2000).

Many studies have been conducted on the psychometric and technical aspects of CAT (for a review, see van der Linden & Glas, 2000). Topics include the construction of the item pool (Kingsbury & Wise, 2000; Lee & Dodd, 2012), the comparison of item selection methods (Finkelman et al., 2014; van der Linden, 2005), and stopping rules (Choi et al., 2011). Previous studies have found that in psychometric and technical terms CAT has many advantages over fixed-item tests (FIT). According to Flens et al. (2016), the number of items in CAT procedures is reduced by 26 to 44 percent, compared with FIT, while the efficiency in testing is actually increased. Other benefits of CAT include improved validity and measurement precision (Linacre,

[1]ELTE Eötvös Loránd University, Budapest, Hungary
[2]Universitas Muhammadiyah Malang, Indonesia

Corresponding Author:
Hanif Akhtar, Doctoral School of Psychology, ELTE Eötvös Loránd University, Izabella utca 46, 1064 Budapest, Hungary.
Email: akhtar.hanif@ppk.elte.hu

2000), while avoiding floor and ceiling effects (Revicki & Cella, 1997). However, CAT also has several drawbacks, including higher development costs, the need for regular item bank maintenance, and complex technical requirements (Tan et al., 2018).

One of the essential issues often neglected is CAT's psychological effect on test-takers. It has been frequently claimed that because in CAT the presented items are matched to test-takers' ability, CAT can be more motivating and less anxiety-inducing than traditional fixed-item tests (Linacre, 2000; Mead & Drasgow, 1993; Wainer, 2000; Weiss, 1982). The reasoning behind this claim is that test-takers with lower ability do not become anxious by items that are too difficult for them, while test-takers with higher ability are bored by items that are too easy for them. Early studies supported this claim: while low-ability examinees answered few items correctly in FIT, they faced easier items in CAT, which made them less discouraged and disengaged than in FIT (Weiss & Betz, 1973). However, high-ability examinees faced more difficult items in CAT, they were more motivated. Furthermore, comparing two versions of a vocabulary test found that both low and high ability students reported higher levels of motivation on the CAT than on the FIT version (Betz, 1977; Betz & Weiss, 1976).

However, the literature on CAT's psychological effects shows mixed results. For example, although Betz and Weiss (1976) and Betz (1977) found that students reported higher level of motivation on the CAT than on the FIT, they also reported higher anxiety in CAT than in FIT. In addition, as questioned by Wise (2014), the test-takers in these early studies were mostly performing a CAT for the first time. Thus, the high level of motivation and anxiety in CAT could be the result of novelty.

Although the accuracy and efficiency of CAT in comparison to FIT are indeed highly relevant from the perspective of test developers, the advantages of CAT are not always perceived by test-takers (Kimura, 2017). Especially, since there is an important characteristic of CAT that makes it inferior for test-takers: Unlike in FIT, they are not allowed to review the test and return to items already administered to change the answer. This feature of CAT does affect the motivational and emotional experience of test-takers (Ortner & Caspers, 2011). Although there are recent developments that aim to implement this feature in CAT (Cui et al., 2018; Han, 2013), in practice, this has not been widely implemented due to its disadvantages such as a complicated test algorithm and increase in testing time (Vispoel et al., 2000).

To our knowledge, currently, there is no systematic review and meta-analysis of the psychological impact of CAT in comparison to FIT. The purpose of this article is to gain a comprehensive understanding of the supposed positive effects of CAT on motivation and anxiety. Motivation refers to test-taking motivation, a particular type of achievement motivation. As the frame of reference, we use the expectancy-value theory (Wigfield & Eccles, 2000), which has two main components: expectancy for success and the perceived value of a task (importance, enjoyment, usefulness of the task, effort). Anxiety refers to state anxiety, defined as a temporary emotional condition elicited from a specific situation (Speilberger, 1972). In this context, state anxiety is anxiety in response to certain testing conditions.

Method

The current study uses the guideline of Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) (Page et al., 2021).

Eligibility Criteria

To be included in this review, studies had to meet the following criteria: (a) original research, (b) written in English, and (c) contained a comparison of state anxiety and/or state motivation (i.e., anxiety and motivation as a reaction of certain testing conditions) between CAT and FIT. The following studies were excluded: (a) oral/poster presentations, (b) studies that did not report original findings, and (c) studies that did not directly compare the effect of CAT versus FIT on state anxiety and motivation. Sample characteristics and test categories were not among the inclusion and exclusion criteria.

Information Sources and Search Strategy

We performed a search on seven databases where we could potentially identify peer-reviewed journal articles as well as gray literature: PsycINFO, PubMed/Medline, Scopus, Google Scholar, ProQuest, EbscoHost Open Dissertation, and Web of Science, for articles published between January 1, 1990 and December 1, 2021, for the following keywords: "computer* adaptive test*," "motivation," "anxiety," with Boolean operators AND and OR—"computer* adaptive test*" AND ("motivation" OR "anxiety") in the title, abstract, or keywords. For the Google Scholar search result, we only extracted the 500 most relevant articles from 3,190 results. The papers referenced in key articles were also reviewed to ensure that no relevant studies were excluded. Duplicate results were removed.

Selection Process

Two reviewers surveyed the title and abstract of each article to select articles that match the inclusion criteria. The shortlisted papers were evaluated for eligibility by the same two reviewers. Any duplicates were deleted from the final pool of papers. When it was necessary, authors of included articles were contacted for Supplementary data.

Data Extraction and Data Items

Two reviewers analyzed the studies, using the following classifications: (a) the psychological aspect investigated in the study (motivation, anxiety, or both), (b) characteristics of participants, (c) the construct measured by the tests, (d) the testing method compared with CAT, (e) the outcome measure, (f) document type, and (g) mean and standard deviation of each group. For the outcome measure, we only extract a measure of state anxiety and/or motivation, that is, motivation and anxiety as a reaction of certain testing conditions. In addition, the specific study design and the nature of the test were also considered in each study. Any disagreements between the reviewers were resolved by consensus. Articles were included only if they featured an independent variable related to the type of testing (i.e., CAT and FIT) and a direct comparison of its effect on state anxiety and/or motivation.

Quality Assessment

In addition to the aspects listed above, studies were also assessed with the Mixed Methods Appraisal Tool-2018 (Hong et al., 2018). Every included study was evaluated first on the basis of (a) the clarity of research questions and (b) whether the collected data are adequate to address the research questions. If the answer was affirmative in both cases, then the included studies were assessed based on study design. Each of the questions was answered with "No," "Yes," or "Cannot tell."

Meta-Analytical Procedures

The 3.3 version of the Comprehensive Meta-Analysis (CMA) software was used for the computation of the individual effect sizes and conducting the analyses (Borenstein et al., 2015). The dependent variable in the present meta-analysis was the standardized mean difference between the CAT and FIT groups on the outcome measures of anxiety and motivation. In consideration of the great variability of sample sizes and different outcome measures in the primary studies, the Hedges' *g* estimate was calculated by using the pooled standard deviations (Hedges, 1983). When more than one appropriate outcome measure was reported in a primary study, the average of these effect sizes was computed. The average effect size and the corresponding 95% confidence interval were calculated using the random-effects model, which incorporate heterogeneity across the included studies (Borenstein et al., 2011). Studies were weighted with the reverse of their variance based on sample size to account for differences (Borenstein et al., 2011). Before calculating the average effect size, individual studies were screened for outlying effect size values, with a standardized residual exceeding ± 3.29 considered as an outlier (Tabachnick & Fidell, 2013). A positive effect size indicated less anxiety or more motivation in the CAT condition compared with the FIT. Instead of Cohen's (1988) classical benchmarks of effect sizes (small = 0.2, medium = 0.5, large = 0.8), benchmarks from social sciences were used (small = 0.05, medium = 0.15, large = 0.20) as suggested by Bakker and colleagues (2019) and Kraft (2020). These benchmarks were further supported by a previous meta-analysis in which a significant positive effect of self-adaptive testing was compared with computerized-adaptive testing with 0.19 Cohen's *d* effect size (Pitkin & Vispoel, 2001).

The heterogeneity of the effect sizes was estimated with the *Q*-statistic and the I^2 estimate, indicating between-study variance caused by systematic differences across primary studies beyond sampling error (Higgins et al., 2021). I^2 values above 75% suggest a substantial relative heterogeneity between primary studies in relation to total variability, which might be explained by factors on the study-level (Higgins et al., 2021). As I^2 informs about the relative percentage of between-study heterogeneity, but not the size of true variance, the absolute random variance was observed as well, referred to as Tau^2 or T^2 (Borenstein et al., 2017).

To address publication bias, gray literature was also included (i.e., theses, conference papers), and the symmetry of Begg's funnel plot and Egger's regression test was examined (Egger et al., 1997). As Sterne and colleagues (2011) suggested, publication bias was tested only for the overall effect, as under 10 studies this test of asymmetry is underpowered. Subgroup analyses were performed to assess different types of CAT tests efficacy compared with FIT, in those cases where there were at least two studies to be included.

Results

The initial search produced 1,208 potential articles, which decreased to 764 after duplicates were removed. The title and abstract of the remaining articles were surveyed according to the inclusion criteria which were met by 27 articles. Finally, after reading the full text of the articles, 11 were included in the study. Thirteen articles were removed because they did not mention any comparison of state motivation and/or state anxiety between CAT and FIT. Three papers were removed because the full-text article was not in English, only the abstract. Figure 1 illustrates the phases of article selection in accordance with PRISMA guidelines.

Characteristics of Included Studies

The characteristics of the included studies are summarized in Table 1. The majority of studies were conducted in western countries: five in the United States (Arvey et al., 1990; Fritts & Marszalek, 2010; Kiskis, 1991; Ling et al.,

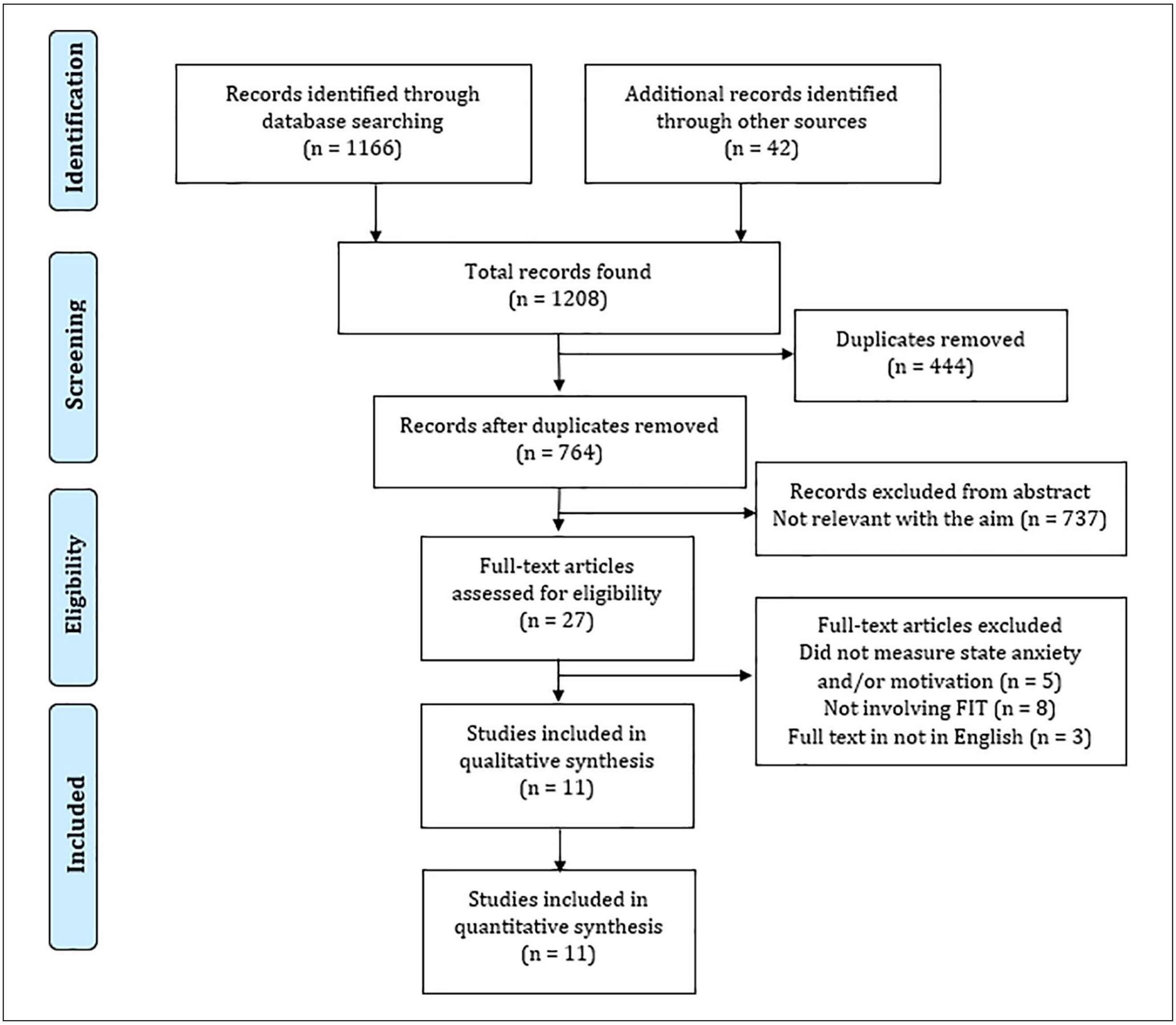

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1166)
Additional records identified through other sources (n = 42)
Total records found (n = 1208)
Duplicates removed (n = 444)
Records after duplicates removed (n = 764)
Records excluded from abstract Not relevant with the aim (n = 737)
Full-text articles assessed for eligibility (n = 27)
Full-text articles excluded Did not measure state anxiety and/or motivation (n = 5) Not involving FIT (n = 8) Full text in not in English (n = 3)
Studies included in qualitative synthesis (n = 11)
Studies included in quantitative synthesis (n = 11)

Figure 1. PRISMA Flowchart of the Current Study.
Note. PRISMA, Preferred Reporting Items for Systematic Review and Meta-Analysis.

2017; Powers, 2001), two in Spain (Olea et al., 2000; Revuelta et al., 2003), one in Germany (Ortner et al., 2014), and one in Australia (Martin & Lazendic, 2018). Only two studies were conducted in a non-Western country: Malaysia (Mohd Ali et al., 2019) and Korea (Kim & McLean, 1995).

The sample size varied considerably in the included studies, ranging from 127 (Kiskis, 1991) to 12,736 (Martin & Lazendic, 2018) participants. All of the studies were conducted in educational settings, except for the study by Arvey et al. (1990) and Kiskis (1991) who conducted their study in organizational setting. All tests measured maximum performance. Four studies compared CAT with Paper-and-Pencil Fixed Item Testing (PPFIT) (Arvey et al., 1990; Fritts & Marszalek, 2010; Kim & McLean, 1995; Powers, 2001), five studies compared CAT with Computer-Based Fixed Item Testing (CBFIT) (Ling et al., 2017; Martin & Lazendic, 2018; Olea et al., 2000; Ortner et al., 2014; Revuelta et al., 2003), and two study compared CAT with both PPFIT and CBFIT (Kiskis, 1991; Mohd Ali et al., 2019).

Quality Assessment of Included Studies

None of the 11 included studies had major problems that endanger their quality. All studies had clearly formulated research questions and reported an appropriate data collection. A few studies did not meet one of the methodological criteria. For example, in Powers' study (Powers, 2001),

Table 1. Summary of Selected Studies Characteristics.

Author(s)	Document type	Country	Psychological aspect	Participants	Construct measured by the test	Testing method to compare	Outcome measure
Kiskis (1991)	Thesis	United States	Anxiety	Applicants at personnel agency (n = 127)	Clerical aptitude	PPFIT, CFIT	STAI, TAI
Kim & McLean (1995)	Conference paper	Korea	Anxiety	College students (n = 208)	Math (algebra)	PPFIT	TAI
Olea et al. (2000)	Journal article	Spain	Anxiety	Undergraduate students (n = 184)	English vocabulary	CFIT	SAS
Powers (2001)	Journal article	United States	Anxiety	GRE Test-takers (n = 1,100)	Verbal reasoning, quantitative reasoning, analytical writing	PPFIT	TAI
Revuelta et al. (2003)	Journal article	Spain	Anxiety	University students (n = 557)	English vocabulary	ECAT, CFIT	SAS
Fritts & Marszalek (2010)	Journal article	United States	Anxiety	Junior high school student (n = 132)	Math and reading ability	PPFIT	STAIC
Mohd Ali et al. (2019)	Journal article	Malaysia	Anxiety	University students (n = 300)	Math (algebra)	CFIT, PPFIT	FTA (SV, CI, and PET)
Arvey et al. (1990)	Journal article	United States	Anxiety, Motivation	Army (n = 535)	Vocational aptitude	PPFIT	TAS (M&S)
Ling et al. (2017)	Journal article	United States	Motivation, Anxiety	Middle school students (n = 789)	Mathematics problem-solving	ECAT, CFIT	QCM (C&I), AQ
Ortner et al. (2014)	Journal article	Germany	Motivation	Secondary school students (n = 174)	Figural reasoning	CFIT	QCM (FF and PS)
Martin & Lazendic (2018)	Journal article	Australia	Motivation	Elementary and secondary school students (n = 12,736)	Numeracy skills	CFIT	MES (PME and NME)

Note. PPFIT = Paper-and-Pencil Fixed-Item Test; CFIT = Computerized Fixed-Item Test; ECAT = Easier Computerized Adaptive Testing; TAS = Test Attitude Survey; M&S = subscale of motivation and comparative anxiety; SAS = State-Anxiety Scale; TAI = Test Anxiety Inventory; STAIC = State-Trait Anxiety Inventory for Children; STAI = State-Trait Anxiety Inventory; FTA = Friedben Test Anxiety Scale; SV, CI, & PET = subscale of Social Views, Cognitive Impairment, and Physical and Emotional Tension; QCM = Questionnaire on Current Motivation; AQ = Anxiety Questionnaire; C&I = Subscale of Challenge and Interest; FF & PS = subscale of Fear of Failure and Probability of Success; MES = Motivation and Engagement Scale; PME & NME = subscale of Positive Motivation and Engagement and Negative Motivation and Engagement.

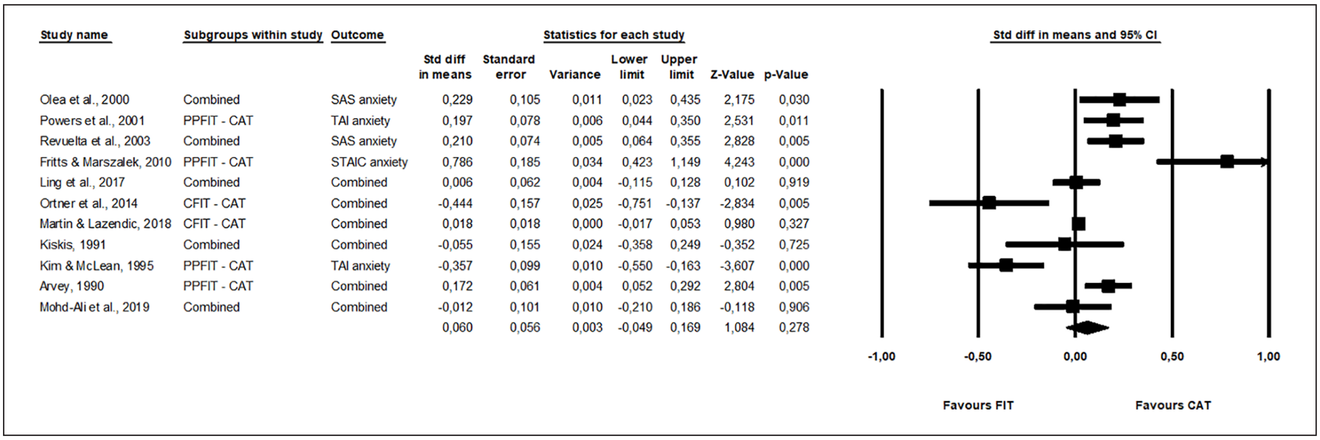
examinees were not randomly assigned to modes of exposure (CAT vs. FIT) but were allowed to self-select themselves into one of the two conditions. In addition, Powers did not control testing mode (computer-based vs. paper-based) and score-reporting (immediately vs. several weeks later) as possible confounders that could affect the result of the study. Another study that did not meet one of the criteria is the only by Fritts and Marszalek (2010) who compared two groups from two different school districts. The testing conditions or test-taker characteristics of the two districts could be different enough to confound the difference in anxiety. The summary of the quality assessment is presented in Table 2.

The included studies used different instruments to measure anxiety and motivation. Anxiety was measured by the following scales: State-Anxiety Scale (Olea et al., 2000; Revuelta et al., 2003), Test Anxiety Inventory (Kim & McLean, 1995; Kiskis, 1991; Powers, 2001), State-Trait Anxiety Inventory (Ling et al., 2017), State-Trait Anxiety Inventory for Children (Fritts & Marszalek, 2010), The Friedben Test Anxiety Scale (Mohd Ali et al., 2019), and Comparative Anxiety subscale of Test Attitude Survey (TAS) (Arvey et al., 1990). Motivation was measured by the Questionnaire on Current Motivation (QCM) (Ling et al., 2017; Ortner et al., 2014), the Short Motivation and Engagement Scale (Martin & Lazendic, 2018), and Motivation subscale of TAS (Arvey et al., 1990).

Some of the studies also reported subscales scores (Arvey et al., 1990; Ling et al., 2017; Martin & Lazendic, 2018; Mohd Ali et al., 2019; Ortner et al., 2014), and some of them reported multiple outcome measures (Kiskis, 1991; Ling et al., 2017). Although two studies (Ling et al., 2017; Ortner et al., 2014) used the QCM as a measure of motivation, they measured different factors; Ling and colleagues measured the "Challenge" and interest' factors and modified the scale to adjust the context of their research, while Ortner and colleagues measured the "Probability of success" and "Fear of failure" factors. Test Anxiety Inventory (TAI) was also administered on one occasion (Fritts & Marszalek, 2010), but we excluded this study in our review since TAI measures trait anxiety (with items such as "I feel very panicky when I take an important test" and it was administered before the achievement tests. In comparison, Powers (2001), Kim and McLean (1995), and Kiskis (1991) have modified the questionnaire TAI to measure test anxiety after taking a test. Some of the studies measured additional constructs, too. For example, Powers (2001), Kiskis (1991), and Fritts and Marszalek (2010) measured computer anxiety. In our review, we only included measures of state anxiety and/or motivation.

Table 2. Risk of Bias Assessment of the Studies on the Effect of CAT on Motivation and Anxiety.

	Screening questions		Methodological quality criteria				
Author(s)	Are the research questions clear?	Do the collected data allow for addressing the research questions?	Are the participants representative of the target population?	Are the measurements appropriate?	Are there complete outcome data?	Are the confounders accounted for in the design and analysis?	During the study period, did the exposure occur as intended?
Kiskis (1991)	Yes	Yes	Yes	Yes	Yes	No	Cannot tell
Kim & McLean (1995)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Olea et al. (2000)	Yes	Yes	Yes	Yes	Yes	Cannot tell	Yes
Powers (2001)	Yes	Yes	Yes	Yes	Yes	No	No
Revuelta et al. (2003)	Yes	Yes	Yes	Yes	Yes	Cannot tell	Yes
Fritts & Marszalek (2010)	Yes	Yes	Yes	Yes	Yes	No	Yes
Ortner et al. (2014)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Arvey et al. (1990)	Yes	Yes	Yes	Yes	Yes	Cannot tell	Yes
Ling et al. (2017)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Martin & Lazendic (2018)	Yes	Yes	Yes	Yes	Yes	No	Yes
Mohd Ali et al. (2019)	Yes	Yes	Yes	Yes	Yes	Cannot tell	Yes



Study name	Subgroups within study	Outcome	Statistics for each study						
			Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Olea et al., 2000	Combined	SAS anxiety	0,229	0,105	0,011	0,023	0,435	2,175	0,030
Powers et al., 2001	PPFIT - CAT	TAI anxiety	0,197	0,078	0,006	0,044	0,350	2,531	0,011
Revuelta et al., 2003	Combined	SAS anxiety	0,210	0,074	0,005	0,064	0,355	2,828	0,005
Fritts & Marszalek, 2010	PPFIT - CAT	STAIC anxiety	0,786	0,185	0,034	0,423	1,149	4,243	0,000
Ling et al., 2017	Combined	Combined	0,006	0,062	0,004	-0,115	0,128	0,102	0,919
Ortner et al., 2014	CFIT - CAT	Combined	-0,444	0,157	0,025	-0,751	-0,137	-2,834	0,005
Martin & Lazendic, 2018	CFIT - CAT	Combined	0,018	0,018	0,000	-0,017	0,053	0,980	0,327
Kiskis, 1991	Combined	Combined	-0,055	0,155	0,024	-0,358	0,249	-0,352	0,725
Kim & McLean, 1995	PPFIT - CAT	TAI anxiety	-0,357	0,099	0,010	-0,550	-0,163	-3,607	0,000
Arvey, 1990	PPFIT - CAT	Combined	0,172	0,061	0,004	0,052	0,292	2,804	0,005
Mohd-Ali et al., 2019	Combined	Combined	-0,012	0,101	0,010	-0,210	0,186	-0,118	0,906
			0,060	0,056	0,003	-0,049	0,169	1,084	0,278

Figure 2. Forest Plot of the Overall Effect of Test Type on Anxiety and Motivation.

Note. This figure demonstrates a forest plot with the individual study effect sizes and the total effect size (Hedges' g) of test type on anxiety and motivation combined. Negative effect size favors the FIT groups (PPFIT and CFIT), positive effect size favors the CAT groups (CAT and ECAT). The total effect is demonstrated in the last row.

Overall Effect of Test Type on Anxiety and Motivation: Meta-Analytical Results

As there were no outlier studies based on the standardized residuals, all 11 studies were included in the meta-analysis of the overall effect of test type on anxiety and motivation. A meta-regression analysis revealed that the year of publication among the included studies had no effect on the overall effect size (coefficient = −0.002, p = .78). The funnel plot showed a symmetrical distribution, which suggested no publication bias (see Figure S1). Similarly, Egger's regression test showed no signs of publication bias (t = 0.51, p = .63). Figure 2 shows the forest plot with a non-significant small effect of test type on overall anxiety and motivation. The overall effect was significantly heterogeneous, with a high proportion of observed variance (84%) reflecting real differences in effect size (see Table 3).

As one of the included studies (Martin & Lazendic, 2018) had a sample size of over 12.000 participants, its relative weight in the overall analysis was twice that of the weight of the smallest sample. For this reason, a sensitivity analysis was performed with the exclusion of the Martin and Lazendic (2018) study (k = 10, $g+$ = .07, SE = .08, 95% confidence interval [CI] = [−0.08, 0.21], p = .37), but still indicating a non-significant small-sized effect.

Subgroup analyses of different comparisons of CAT, PPFIT, and CFIT were non-significant, except for ECAT's overall effect on motivation and anxiety in contrast to PPFIT and CFIT, showing a large positive effect (see Table 3.)

Table 3. Effects of Testing Type on Anxiety and Motivation.

	Effects based on standardized mean differences and heterogeneity								
Effects	*k*	Mean effect size (g^+)	95% CI	*p*	*SE*	*Q* value	*p*	I^2 (%)	T^2
Overall effect	11	0.06	[−0.05, 0.17]	.28	0.06	61.46	.001	84	0.02
Effect favors CAT to PPFIT and CFIT	11	0.04	[−0.09, 0.16]	.56	0.06	67.37	.001	85	0.03
Effect favors CAT to PPFIT	6	0.11	[−0.14, 0.35]	.39	0.12	39.26	.001	87	0.07
Effect favors CAT to CFIT	7	−0.02	[−0.16, 0.12]	.79	0.07	24.25	.001	75	0.02
Effect favors ECAT to PPFIT and CFIT	2	0.22	[0.09, 0.36]	.001	0.07	0.08	.77	0	0.01
Anxiety	9	0.09	[−0.06, 0.23]	.23	0.07	46.37	.001	83	0.04
Effect favors CAT to PPFIT and CFIT	9	0.06	[−0.11, 0.23]	.49	0.09	57.43	.001	86	0.06
Effect favors CAT to PPFIT	6	0.08	[−0.16, 0.31]	.52	0.12	36.82	.001	86	0.07
Effect favors CAT to CFIT	5	0.02	[−0.22, 0.26]	.86	0.12	20.37	.001	80	0.06
Effect favors ECAT to PPFIT and CFIT	2	0.22	[0.09, 0.35]	.001	0.07	0.05	.82	0	0.01
Motivation	4	0.03	[−0.15, 0.21]	.75	0.09	31.67	.001	91	0.03
Effect favors CAT to PPFIT and CFIT	4	−0.03	[−0.25, 0.19]	.78	0.11	36.48	.001	92	0.04
Effect favors CAT to CFIT	3	−0.15	[−0.38, 0.07]	.18	0.12	12.28	.002	84	0.03

Note. k = number of included studies; g^+ = Hedges' *g* effect size; 95% CI = 95% confidence interval; *p* = significance value; *Q* = Cochran's *Q* value to test heterogeneity; I^2 = percentage of relative variance across studies due to heterogeneity; T^2 = absolute between-study variance; CAT = computerized adaptive testing; ECAT = Easier Computerized Adaptive Testing; PPFIT = Paper-and-Pencil Fixed Item Testing; CFIT = Computerized Fixed Item Testing.

Effect of CAT on Anxiety

Four (Fritts & Marszalek, 2010; Ling et al., 2017; Mohd Ali et al., 2019; Powers, 2001) of the nine articles that discuss anxiety found significantly lower levels of reported anxiety when taking a CAT. Fritts and Marszalek (2010) compared state anxiety of junior high school students after taking a standardized achievement test. The result of the analysis showed that examinees who took a traditional test had a higher mean state anxiety score than examinees who took the CAT, after controlling for computer anxiety and test anxiety. Powers (2001) also compared examinees' anxiety after they took the Graduate Record Examination (GRE) Test—albeit several days after actually taking the test—and found that the Paper Based Test (PBT) sample reported higher anxiety levels than the CAT sample. The same result was also found by Ali and colleagues (Mohd Ali et al., 2019), who compared college students' anxiety when taking mathematics tests and found that CAT reduced examinees' anxiety in comparison to those who took a traditional fixed-item test (PBT and CBT). The same effect of CAT on anxiety was also found by Ling and colleagues (2017). However, they used two types of CAT: Easier CAT (ECAT) and regular CAT. The ECAT was a version of CAT in which items were chosen at a lower difficulty level than the examinee's estimated ability, thus increasing the probability of arriving at a correct answer from the 50% that is regularly applied in a CAT. They compared middle school students' state anxiety after taking mathematics problem-solving tests and found that ECAT resulted in lower anxiety than either regular CAT or CFIT.

Five of the nine studies (Arvey et al., 1990; Kim & McLean, 1995; Kiskis, 1991; Olea et al., 2000; Revuelta et al., 2003) did not find a statistically significant effect of test condition on anxiety. The goal of the study by Olea and colleagues was to examine the effect of being able to review and change previous answers on computerized tests, both fixed and adaptive, they also compared participants' state anxiety before and after taking an English vocabulary test. A similar study was conducted by Revuelta and colleagues (2003). Their main goal was to investigate the effect of item selection and the ability to review previous items on computerized testing. However, they also compared participants' state anxiety among three types of tests: CAT, ECAT, and CFIT. Arvey et al. (1990) compared anxiety of Armies after taking CAT and FIT version of The Armed Service Vocational Aptitude Battery (ASVAB), while Kiskis (1991) compared anxiety of applicants at personnel agency after taking CAT and FIT version of clerical aptitude test.

Meta-Analytical Results

As shown in Table 3, a nonsignificant small effect of testing type on anxiety was found. The effect was heterogeneous, with 83% of the observed variance reflecting differences in effect size.

Subgroup analyses of different comparisons of CAT, PPFIT, and CFIT were nonsignificant, except for ECAT's overall effect on anxiety in contrast to PPFIT and CFIT, indicating a large positive effect (see Table 3.)

Effect of CAT on Motivation

Two of the four articles reported a positive effect of CAT on motivation (Arvey et al., 1990; Ling et al., 2017). Arvey et al. (1990) reported that the CAT version of the ASVAB had significantly higher scores on the Motivation factors compared with the paper-and-pencil version of the ASVAB. In addition, Ling and colleagues (2017) compared three types of tests: ECAT, regular CAT, and CFIT and found that ECAT resulted in higher motivation than regular CAT or CFIT. However, they did not find any significant difference of motivation between regular CAT and CFIT.

Another study compared test-relevant motivation and engagement in elementary and secondary school students who completed a numeracy test and reported the lack of a statistically significant effect of test condition on motivation (Martin & Lazendic, 2018). Finally, one of the papers even reported a negative effect of CAT on motivation in secondary school students (Ortner et al., 2014). During a break in the testing session, state motivation was measured, and "Fear of failure" was higher in the CAT condition than in the CFIT condition. Moreover, the "probability of success" in the CAT condition was lower than in the CFIT condition. These results might explain why students found CAT to me more motivating than CFIT.

Meta-Analytical Results

As shown in Table 3, there was a nonsignificant small effect of testing type on motivation. The effect was heterogeneous with 91% of the observed variance reflecting differences in effect size.

As the Martin and Lazendic (2018) study, with a sample size of over 12.000 participants and a relative weight twice of the weight of the smallest study in the subgroup analysis, a sensitivity analysis was performed with the exclusion of this study ($k = 3$, $g+ = .005$, $SE = .17$, 95% CI = [−0.32, 0.33], $p = .98$), but still indicating a non-significant small sized effect.

Subgroup analyses about different comparisons of CAT, PPFIT, and CFIT were non-significant (see Table 3), although ECAT type of testing was not compared with FIT types of tests as there was only one study measuring this.

Discussion

This review examined the effect of CAT on motivation and anxiety in comparison to traditional FIT, based on 11 studies. The general result of our review and meta-analysis suggested no significant effect of test type on anxiety and motivation when comparing CAT with FIT. This is in contrast with the claims articulated in early work on CAT (Betz, 1977; Betz & Weiss, 1976; Linacre, 2000; Wainer, 2000; Weiss, 1982; Weiss & Betz, 1973).

Only two of the four studies on motivation and four of nine studies on anxiety in our review supported the benefits of CAT, while one of them showed the opposite result: a decrease in motivation under CAT. It should be also noted that the single study which demonstrated a positive effect of CAT on motivation and anxiety (Ling et al., 2017) compared two types of CAT, easier CAT (ECAT) and regular CAT. They found that only ECAT, but not traditional CAT, resulted in higher motivation and lower anxiety than regular CFIT.

It is possible that there are methodological reasons for the null findings. For example, in the study of Ortner and colleagues (2014), test-takers were not given specific information about how CAT works. That such information might be relevant is highlighted in an earlier study (Ortner & Caspers, 2011) that also found a high level of anxiety in CAT, but only when no explanation was provided to participants. Another possibility is that participants are uncomfortable with certain features in CAT, such as the inability to review or skip items (Tonidandel et al., 2002; Tonidandel & Quiñones, 2000). The difference between low-stakes versus high-stakes testing situations could also affect motivation and anxiety. For example, Revuelta and colleagues noted that a lack of an effect of test type on anxiety may be due to the floor effect caused by the low-stakes nature of the test (Revuelta et al., 2003). However, in high-stakes testing (e.g., in the GRE test), Powers found that those who took PBT reported more anxiety than those who took CAT (Powers, 2001). Uncontrolled confounders were also found in few studies, such as different school districts (Fritts & Marszalek, 2010) or other pre-existing differences (Powers, 2001).

In addition, several of the reviewed studies also discussed the different conditions of CAT that could affect motivation and anxiety. In our analysis, using ECAT had significant large effect on anxiety in comparison with FIT. It was in line with previous studies (Häusler & Sommer, 2008; Tonidandel & Quiñones, 2000) that found respondents' reactions to be more favorable under easier computerized adaptive tests. However, using easier items is not optimal from the perspective of measurement efficiency (Bergstrom et al., 1992; Häusler & Sommer, 2008). For example, it takes 100 items to reach a SEM of .20 if the probability of a correct response is 50%, 104 items if 60%, and 119 items if 70% (Bergstrom et al., 1992). However, the increase in test length did not lead to an increase in test duration (Häusler & Sommer, 2008).

Another condition that could lower examinees' level of state-anxiety is allowing them to review previously administered items and change their responses (Olea et al., 2000; Revuelta et al., 2003). However, from the perspective of test developers permitting item review is difficult, since the test algorithm has to be more complicated and testing time typically increases by 37% to 61% (Vispoel et al., 2000).

Furthermore, the specific procedures employed by the reviewed studies also provide valuable information about the psychological aspects of using CAT. For example, Olea and colleagues (2000) suggested that providing detailed, item-level feedback on performance after the exam leads to decreased state anxiety and an increased ability estimate level. Future investigation in this topic is needed.

The ability level of the examinees might also mediate results. In our review, only three studies investigated the relationship between performance, testing mode, and psychological effects (Ling et al., 2017; Ortner et al., 2014; Powers, 2001). Ling and colleagues (2017) reported that higher ability examinees tended to report less anxiety and less engagement for each mode of testing (CAT, ECAT, and FIT). However, under the ECAT condition, lower ability examinees reported less anxiety and more engagement than in regular CAT and FIT conditions. A similar result was found by Powers (2001): the relationship between performance and anxiety was similar for each mode of testing (CAT and FIT). Yet a different result was reported by Ortner and colleagues (2014): motivation was equal for high- and low-performance examinees in the CAT condition, but in the FIT condition, high-performance examinees experienced a higher motivation. Evidence for the interaction between ability and mode of testing is still inconclusive, and thus future research in this area is required. Unfortunately, due to the lack of available data, we were unable to carry out a quantitative analysis of this issue.

This review has several limitations. First, it only considered studies that contained a comparison of motivation and/or anxiety between CAT and FIT, but not a comparison within CAT conditions, such as the ones carried out by Häusler and Sommer (2008) as well as Tonidandel and colleagues (2002), who compared different item selection methods and their impact on examinee's motivation.

Second, the number of studies included for the meta-analysis was small. However, as Davey and colleagues (2011) reported the average meta-analysis in some fields includes a median of three studies. Third, our review only included English-language studies. Fourth, several of the reviewed studies did not control for possible confounder variables such as trait anxiety, computer anxiety, test-taker's ability, and testing context (low- and high- stakes).

Based on the results of our study, there are several suggestions for test-developers as well as researchers of CAT. First, there are several conditions under which CAT could affect test-takers' motivation and anxiety, such as the opportunity to review items or using easier items than the examinee's estimated ability. Thus, CAT developers might want to consider modifying the CAT algorithm to optimize the experience from a psychological perspective. Second, it has been suggested that test-takers' motivation and anxiety affect achievement (e.g., Robbins et al., 2004). To increase fairness in CAT, future research should explore the relationship between motivation, anxiety, and test performance and explore features of testing that avoid negative psychological effects.

Conclusion

We reviewed evidence of the effect of CAT on test-takers' motivation and anxiety, in comparison to FIT. In conclusion, our review suggests that overall, there is no effect of mode of testing on anxiety and motivation. However, when comparing ECAT with FIT testing, samples tested with ECAT showed less anxiety. In addition, certain modifications in CAT administration such as presenting easier items can provide positive psychological effects for test-takers. These modifications, however, are less favorable for test developers and psychometricians because they decrease measurement efficiency. Therefore, practical considerations should be made to maximize the trade-off between test-takers' psychological experience and measurement efficiency.

Moreover, these results can have implications for test design. CAT and FIT have different procedures for item selection and scoring, and it is possible that an interaction exists between psychological effects of the type of test and the test-takers' level of ability. That is, in regular CAT, test-takers are usually faced with items that have a success probability of 50%, consequently, they will complete about half of the items correct, regardless of their ability. In FIT, however, the number of correct answers depends on test-takers' ability; higher ability examinees have more items correct. Test-takers' emotional reaction might be affected by their perception of how many items they answered correctly. In particular, their experience with CAT might be drastically different from what they are used to in FIT: High-ability examinees get fewer items right than what they are used to, while lower ability examinees get the impression that they score better than usual. In relation to test fairness, future research could investigate what information about adaptive testing should be provided to avoid such negative psychological effects.

Author Contributions H.A. and S. performed the database search, reviewed the selected studies, and summarized the results. B.V. performed the meta-analysis and summarized the results. H.A. provided a first draft of the paper, which was finalized by K.K. All authors saw and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Kristof Kovacs received funding by the National Research, Development and Innovation Office of Hungary: Grant FK-21-138971, by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and by the ÚNKP-21-5 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and

Innovation Fund. Boglarka Vekety received funding by the ÚNKP-21-4 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund.

Data Availability Statement

The data sets analyzed in this study can be found in the OSF Repository (https://osf.io/qm9sj/?view_only=b02d83c0f501440d957806665bbc9e13).

ORCID iD

Hanif Akhtar https://orcid.org/0000-0002-1388-7347

Supplemental Material

Supplemental material for this article is available online.

References

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, *43*(4), 695–716. https://doi.org/10.1111/j.1744–6570.1990.tb00679.x

Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., & Van Dooren, W. (2019). Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics*, *102*(1), 1–8. https://doi.org/10.1007/s10649-019-09908-4

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the difficulty in computer adaptive testing. *Applied Measurement in Education*, *5*, 137–149.

Betz, N. E. (1977). Effects of immediate knowledge of results and adaptive testing on ability test performance. *Applied Psychological Measurement*, *1*, 259–266.

Betz, N. E., & Weiss, D. J. (1976). *Psychological effects of immediate knowledge of results and adaptive ability testing* (Research Report, 76-4). University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. Wiley.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2015). Regression in meta-analysis. *Comprehensive Meta Analysis*. https://www.meta-analysis.com/pages/cma_manual.php

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *71*(1), 37–53. https://doi.org/10.1177/0013164410387338

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Cui, Z., Liu, C., He, Y., & Chen, H. (2018). Evaluation of a new method for providing full review opportunities in computerized adaptive testing—Computerized adaptive testing with salt. *Journal of Educational Measurement*, *55*(4), 582–594. https://doi.org/10.1111/jedm.12193

Davey, J., Turner, R. M., Clarke, M. J., & Higgins, J. P. (2011). Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: A cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, *11*(1), 1–11. https://doi.org/10.1186/1471-2288-11-160

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *The BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Finkelman, M. D., Kim, W., Weissman, A., Cook, R. J., Veldkamp, B. P., Barnard, J., van der Linden, W. J., Ramón Barrada, J., Mead, A. D., Becker, K. A., Reckase, M. D., Dodd, B. G., Riley, B., Eggen, T., Walter, O. B., Frey, A., Wang, W.-C., Han, K. T., & Wise, S. L. (2014). Cognitive diagnostic models and computerized adaptive testing: Two new item-selection methods that incorporate response times. *Journal of Computerized Adaptive Testing*, *2*(3), 59–76. https://doi.org/10.7333/1412-0204059

Flens, G., Smits, N., Carlier, I., van Hemert, A. M., & de Beurs, E. (2016). Simulating computer adaptive testing with the mood and anxiety symptom questionnaire. *Psychological Assessment*, *28*(8), 953–962. https://doi.org/10.1037/pas0000240

Fritts, B. E., & Marszalek, J. M. (2010). Computerized adaptive testing, anxiety levels, and gender differences. *Social Psychology of Education*, *13*(3), 441–458. https://doi.org/10.1007/s11218-010-9113-3

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*(4), 347–360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x

Han, K. T. (2013). Item pocket method to allow response review and change in computerized adaptive testing. *Applied Psychological Measurement*, *37*(4), 259–275. https://doi.org/10.1177/0146621612473638

Häusler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science Quarterly*, *50*(1), 75–87.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*(2), 388–395. https://doi.org/10.1037/0033-2909.93.2.388

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M., & Welch, V. (Eds.). (2021). *Cochrane handbook for systematic reviews of interventions* (version 6). Cochrane. https://www.training.cochrane.org/handbook

Hong, Q., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). *Mixed Methods Appraisal Tool (MMAT), version 2018*. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada.

Kim, J., & McLean, J. E. (1995, April). The influence of examinee test-taking behavior motivation in computerized adaptive testing [Paper presentation]. Annual meeting of the National

Council on Measurement in Education, San Francisco, CA, United States.

Kimura, T. (2017). The impacts of computer adaptive testing from different perspectives. *Journal of Educational Evaluation for Health Professions*, *14*, 12. https://doi.org/10.3352/jeehp.2017.14.12

Kingsbury, G., & Wise, S. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica: Revista de Metodología y Psicología Experimental*, *21*(1), 135–156.

Kiskis, S. (1991). Effects of test administrations on general, test, and computer anxiety, and efficacy measures. *Theses Digitization Project*, *579*. https://scholarworks.lib.csusb.edu/etd-project/579

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Lee, H. Y., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for CAT using the partial credit model. *Educational and Psychological Measurement*, *72*(1), 159–175. https://doi.org/10.1177/0013164411411296

Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come* (No. 69, p. 58). MESA memorandum. https://www.cehd.umn.edu/EdPsych/C-Bas-R/Docs/Linacre2000_CAT.pdf

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Lunz, M. E., Bergstrom, B. A., & Gershon, R. C. (1994). Computer adaptive testing. *International Journal of Educational Research*, *21*(6), 623–634. https://doi.org/10.1016/0883–0355(94)90015-9

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27–45. https://doi.org/10.1037/edu0000205

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests : A equivalence of computerized and paper-and-pencil cognitive ability tests : A meta-analysis. *Psychological Bulletin*, *114*(3), 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Mohd Ali, S., Norfarah, N., Ilya Syazwani, J. I., & Mohd Erfy, I. (2019). The effect of computerized-adaptive test on reducing anxiety towards math test for polytechnic students. *Journal of Technical Education and Training*, *11*(4), 27–35. https://doi.org/10.30880/jtet.2019.11.04.004

Olea, J., Revuelta, J., Ximénez, M., & Abad, F. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive test. *Psicológica: Revista de Metodología y Psicología Experimental*, *21*(1), 157–174.

Ortner, T. M., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*, *27*(3), 157–163. https://doi.org/10.1027/1015-5759/a000062

Ortner, T. M., Weißkopf, E., & Koch, T. (2014). I will probably fail: Higher ability students' motivational experiences during adaptive achievement testing. *European Journal of Psychological Assessment*, *30*(1), 48–56. https://doi.org/10.1027/1015-5759/a000168

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ*, *372*, Article n71. https://doi.org/10.1136/bmj.n71

Pitkin, A. K., & Vispoel, W. P. (2001). Differences between self-adapted and computerized adaptive tests: A meta-analysis. *Journal of Educational Measurement*, *38*(3), 235–247. https://doi.org/10.1111/j.1745–3984.2001.tb01125.x

Powers, D. E. (2001). Test anxiety and test performance: Comparing paper-based and computer-adaptive versions of the graduate record examinations (GRE©) general test. *Journal of Educational Computing Research*, *24*(3), 249–273. https://doi.org/10.2190/680W-66CR-QRP7-CL1F

Revicki, D. A., & Cella, D. F. (1997). Health status assessment for the twenty-first century: Item response theory, item banking and computer adaptive testing. *Quality of Life Research*, *6*(6), 595–600. https://doi.org/10.1023/A:1018420418455

Revuelta, J., Ximénez, M. C., & Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. *Educational and Psychological Measurement*, *63*(5), 791–808. https://doi.org/10.1177/0013164403251282

Robbins, S. B., Le, H., Davis, D., Lauver, K., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, *130*(2), 261–288. https://doi.org/10.1037/0033-2909.130.2.261

Speilberger, C. D. (Ed.). (1972). *Anxiety: Current trends in theory and research* (Vols. 1–2). Academic Press.

Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *The BMJ*, *343*, Article d4002. https://doi.org/10.1136/bmj.d4002

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Tan, Q., Cai, Y., Li, Q., Zhang, Y., & Tu, D. (2018). Development and validation of an item bank for depression screening in the Chinese population using computer adaptive testing: A simulation study. *Frontiers in Psychology*, *9*(JUL), Article 1225. https://doi.org/10.3389/fpsyg.2018.01225

Tonidandel, S., & Quiñones, M. A. (2000). Psychological reactions to adaptive testing. *International Journal of Selection and Assessment*, *8*(1), 7–15. https://doi.org/10.1111/1468–2389.00126

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, *87*(2), 320–332. https://doi.org/10.1037/0021–9010.87.2.320

van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement*, *42*(3), 283–302. https://doi.org/10.1111/j.1745-3984.2005.00015.x

van der Linden, W. J., & Glas, G. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Springer. https://doi.org/10.1007/0–306-47531–6

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1–25). Springer. https://doi.org/10.1007/0-306-47531-6_1

Vispoel, W. P., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement*, *37*(1), 21–38. https://doi.org/10.1111/J.1745-3984.2000.TB01074.X

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum.

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large scale computerized testing. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 271–299). Lawrence Erlbaum.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473–492.

Weiss, D. J., & Betz, N. E. (1973). *Ability measurement: Conventional or adaptive?* (Research Report 73-1). Department of Psychology, Psychometric Methods Program.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68–81.

Wise, S. L. (2014, January). The utility of adaptive testing in addressing the problem of unmotivated examinees the utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, *2*(1). https://doi.org/10.7333/1401-0201001