Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test?

Applied Psychological Measurement

2017, Vol. 41(7) 495–511

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621617707556

journals.sagepub.com/home/apm

Guangming Ling¹, Yigal Attali¹,
Bridgid Finn¹, and Elizabeth A. Stone¹

Abstract

Computer adaptive tests provide important measurement advantages over traditional fixed-item tests, but research on the psychological reactions of test takers to adaptive tests is lacking. In particular, it has been suggested that test-taker engagement, and possibly test performance as a consequence, could benefit from the control that adaptive tests have on the number of test items examinees answer correctly. However, previous research on this issue found little support for this possibility. This study expands on previous research by examining this issue in the context of a mathematical ability assessment and by considering the possible effect of immediate feedback of response correctness on test engagement, test anxiety, time on task, and test performance. Middle school students completed a mathematics assessment under one of three test type conditions (fixed, adaptive, or easier adaptive) and either with or without immediate feedback about the correctness of responses. Results showed little evidence for test type effects. The easier adaptive test resulted in higher engagement and lower anxiety than either the adaptive or fixed-item tests; however, no significant differences in performance were found across test types, although performance was significantly higher across all test types when students received immediate feedback. In addition, these effects were not related to ability level, as measured by the state assessment achievement levels. The possibility that test experiences in adaptive tests may not in practice be significantly different than in fixed-item tests is raised and discussed to explain the results of this and previous studies.

Keywords

adaptive testing, feedback, motivation, anxiety, effort, low-stakes assessments

Computer adaptive tests (CATs) rank among the most important developments in psychological assessment (Wainer, 2000). In a CAT, an examinee's ability is iteratively estimated after each test item is answered, and subsequent items are selected on the basis of the current ability estimate (e.g., Lord, 1980). As in traditional testing, ability estimates in a CAT depend on the correctness of the examinee's responses. However, in a CAT they also depend on the difficulty of the specific items selected for the examinee. The purpose of this adaptive item selection process is to "tailor" test items to the current ability estimate of the examinee so that items provide

¹Educational Testing Service, Princeton, NJ, USA

Corresponding Author:

Guangming Ling, Educational Testing Service, 660 Rosedale Road, MS 07-R, Princeton, NJ 08541, USA.

Email: gling@ets.org

maximum information about the continued estimation of examinee ability. In particular, through this process examinees are not asked to answer items that are expected to be either very easy or very difficult for them. When items are either very easy or very difficult for the test taker to answer, the uncertainty about whether such an item can be answered correctly is low, and therefore, these items provide relatively little information about the examinee's ability. In contrast, an item for which the examinee has a 50% expected probability of answering correctly is most informative in improving his or her current ability estimate.

This principle of item selection has resulted in a significant theoretical psychometric advantage for CATs—increased efficiency in testing. Because items are tailored to each examinee and provide more information about their ability level, measurement is expected to be more accurate in a CAT than in the usual framework of fixed-item tests (FIT). Alternatively, tests under the CAT framework can be shorter than a FIT while maintaining the same accuracy of measurement (Weiss, 1982; Weiss & Kingsbury, 1984). Moreover, in principle the same level of measurement accuracy could be obtained for all examinees across the ability range, in contrast to FITs, which usually result in lower measurement accuracy for low- and high-ability examinees because FIT items target middle ability examinees (to achieve higher overall measurement accuracy).

The psychometric and technical aspects of CAT have been the subject of many investigations (for a review, see Van der Linden & Glas, 2000). As an example, item selection-related issues have been the subject of continuing technical developments, as real-life implementations of CATs have required more complex considerations of content coverage and item pool exposure in addition to the principle of information maximization (Chang & Ying, 1999; Cheng & Chang, 2009). However, few studies have dealt with the psychological effects that CATs might have on examinees and how these effects could impact test performance (Ortner, Weißkopf, & Koch, 2014). In particular, because in a CAT all examinees answer around 50% of the items correctly, the test experience of many examinees could be very different than in a FIT.

These disparate test experiences could have motivational consequences, which in turn could influence test performance. Based on Atkinson's (1957) theory of achievement motivation, expectancy-value theory (Wigfield & Eccles, 2000) is one of the most important conceptions of achievement motivation, including test performance. The main components of the theory, expectancy for success and the perceived value of a task, are assumed to affect achievement behavior, performance, as well as effort and persistence. The expectancies refer to the students' beliefs of how well they will perform. The value component consists of four distinct aspects: attainment value (importance), intrinsic value (enjoyment), utility value (usefulness of the task), and cost (effort). The value aspects are assumed to explain performance-related decisions based on students' beliefs about how they might benefit from a task.

The expectancy-value model is also the framework most commonly used to conceptualize test-taking motivation (e.g., Penk & Schipolowski, 2015), which is a particular type of achievement motivation. Theories of motivation have also acknowledged that in specific situations, task characteristics (including test features) play an important role in determining motivation and behavior (Vollmeyer & Rheinberg, 2006). Two types of arguments were made regarding the possible beneficial effects that CATs might have on examinee motivation during the test, which could in turn boost performance. These arguments, and the attempts to empirically test them, are reviewed next.

The first argument was made early on by Weiss and Betz (1973), who suggested that the distinctive feature of CATs might have a beneficial effect on examinee motivation and engagement. In particular, they argued that, since low-ability examinees will experience an easier test (compared with a FIT), they may become less discouraged or disengaged during a CAT because in a FIT, these lower ability examinees answer few items correctly. Furthermore, as high-ability

examinees will experience a more difficult test than they are used to in a FIT, they may become less disengaged due to boredom. These possible motivational benefits could in turn result in higher performance. This possibility could have important practical implications, as performance of students in low-stakes testing situations has been a concern and focus of recent research (Wise & DeMars, 2010).

However, there is little evidence to support this suggestion (see also Wise, 2014). Although there exists a relatively large body of research focusing on comparability of performance on paper-and-pencil and computerized tests (summarized in several meta-analyses, see Bergstrom, 1992; Mead & Drasgow, 1993; Powers, 2001; Wang, Jiao, Young, Brooks, & Olson, 2007), only a few studies involved CATs. However, comparisons of CATs with paper-and-pencil tests confound two possible effects (computer-administered testing and adaptivity) and therefore have limited value in studying the question of how adaptivity influences motivation and performance.

The only direct comparison of performance on a CAT and FIT was conducted by Betz and Weiss (1976) and Betz (1977), who administered a vocabulary test as either a FIT or a CAT. They found that low-ability students reported significantly higher levels of motivation on the CAT than on the FIT, high-ability students reported similar motivation levels between the FIT and CAT, and both low- and high-ability students reported significantly higher levels of anxiety on the CAT than on the FIT. However, they did not find a significant effect of test type on test performance. In addition, a few other studies compared self-reported motivation under CAT and FIT and found higher motivation under CAT (Arvey, Strickland, Drauden, & Martin, 1990; Pine, Church, Gialluca, & Weiss, 1979) but did not examine whether these motivational benefits translated to a performance boost.

The second argument made regarding the possible beneficial effect that CATs might have on examinee motivation was first posed by Bergstrom, Lunz, and Gershon (1992), who suggested that an easier CAT, although not optimal from the perspective of measurement efficiency, might “ease examinee anxiety or bolster esteem.” Two studies by Bergstrom and Lunz (Bergstrom et al., 1992; Lunz & Bergstrom, 1994) manipulated the difficulty of a medical technology certification CAT (with 50%, 60%, and 70% targeted difficulty). Although the easier tests were slightly less reliable, which was expected because less than optimal item selection criteria were used, performance on the various adaptive tests was unaffected by test difficulty (affective reactions to the tests were not measured in these studies). Similarly, two later studies manipulated test difficulty of a vocabulary CAT but did not find an effect on either ability estimates or posttest anxiety (Häusler & Sommer, 2008; Ponsoda, Olea, Rodriguez, & Revuelta, 1999).

In summary, two types of arguments were made regarding the possible beneficial effects that CATs might have on examinee motivation on the test, which could in turn boost performance. One argument was that, compared with a FIT, a CAT could lower anxiety for low-ability examinees and lower boredom for high-ability examinees, thereby increasing motivation for these groups of examinees (Weiss & Betz, 1973). The other argument was that, compared with a standard CAT with 50% expected success rate, an easier CAT could lower anxiety for all examinees (Bergstrom et al., 1992). However, research did not find support for these arguments.

Motivation for Current Study

Three aspects of previous research on this topic are notable and serve as motivation for the current study. First, the research focused on tests of declarative knowledge (knowledge of medical technology and vocabulary). Answering items in these tests is based on retrieval of facts from memory. As such, these tests are less cognitively demanding, and require less effort, than more

cognitively taxing tasks such as mathematical problem solving. It is reasonable to assume that test conditions that engender higher test-taking motivation would have a more pronounced beneficial effect with more cognitively taxing items because of the extra effort required to answer these items. Consistent with this prediction, Wolf, Smith, and Birnbaum (1995) found that students with lower test-taking motivation performed reasonably well on items that were not mentally taxing, but not on those that were mentally taxing. Similarly, DeMars (2000) found that constructed response items (e.g., essays) were more vulnerable to diminished effort than multiple-choice items (see Martinez, 1999, for a discussion of the differences in cognition required by the different item formats).

A second notable aspect of most of the previous research is the lack of immediate feedback to examinees. Providing feedback regarding task performance is one of the most frequently applied of all psychological interventions and has a major role in learning and instruction (see Kluger & DeNisi, 1996, for a review). Feedback helps learners determine performance expectations, judge their level of understanding, and become aware of misconceptions. Yet, feedback has had almost no place in assessment (Attali & Powers, 2010). This is particularly problematic in the context of the potential effect of adaptive testing on examinee motivation and performance. The lack of feedback may undermine any beneficial effects of adaptivity because examinees would not have definite knowledge about their performance during the test, and instead would have to rely on metacognitive estimates of response correctness. In theory, examinees taking a CAT without feedback could use their own estimates of item difficulty to gauge whether or not they answered the previous item correctly. However, even for experts, the task of estimating item difficulty is very hard (Impara & Plake, 1998). Betz (1977) was the only study that provided immediate feedback (termed knowledge of results) during the CAT and FIT, although without making the argument that test type effects could be stronger with than without feedback. Betz (1977) found a significant positive effect for immediate feedback of results but no interaction between presence of feedback and test type.

Last, the relatively small average group sample sizes of the research reviewed should be noted. With the exception of Lunz and Bergstrom (1994) with 215 examinees per group, all other studies had between 45 and 75 examinees per condition. Taking into account that previous research on test format effects (Bergstrom, 1992; Mead & Drasgow, 1993; Wang et al., 2007) found generally small effects, larger sample sizes might be needed to ensure adequate statistical power.

Current Study

In view of these considerations and previous research on the topic, two general questions are examined:

1. Are reactions to an adaptive test more favorable than a FIT?
2. Are reactions to an easier adaptive test more favorable than a regular adaptive test?

These questions are examined in the context of a mathematics problem-solving test and with a relatively large sample size (789 students). The test is administered either with or without immediate feedback of results to consider the possible mediating effect of feedback. The possible mediating effect of ability (as measured independently through the state assessment of mathematics) is also examined.

To examine motivational effects (in addition to test performance), two types of posttest self-reports (engagement and anxiety) and a behavioral index of effort and persistence (time on task) are measured.

The posttest self-reports were influenced by Vollmeyer and Rheinberg's (2006) model of current achievement motivation that differentiates four distinct factors. These factors are (a) anxiety, (b) challenge, (c) interest, and (d) probability of success. Although this definition takes individual differences in task preferences into account, it is conceptually defined as motivation before completing a task ("initial motivation"). As the purpose was to measure motivational reactions to the task after it was completed, the probability of success factor was not relevant to the investigation. Positive affective reactions to the test were conceptualized as student engagement, comprising of challenge and interest in the test just completed. Interest is related to a positive evaluation of a task and its appeal. Challenge concerns the degree to which the task is accepted as relevant. Negative affective reaction is captured by anxiety during the task and can be interpreted as fear of failure in the achievement situation (Vollmeyer & Rheinberg, 2006).

In addition to self-report judgments, effort and persistence as measured by time on task is commonly used as a measure of motivation (Pintrich, 2003; Schunk, Meece, & Pintrich, 2014). In this study, individual participants' response times on each test item were interpreted in comparison to response times from an independent sample obtained through an earlier study with the same set of items.

Method

Participants

A suburban middle school in Florida was recruited for the study. The school agreed to have all students complete the study. Compensation of US\$10 was provided to the school for each student completing the study. No compensation was provided to students. A total of 789 students across Grades 6 to 8 participated in this study as part of a regular 45-min class in a computer lab of the school. The sample consisted of 49% female students; 42% White students, 36% African American students, 10% Hispanic students, 8% multiracial students, and 4% Asian American students.

Instruments

Mathematics assessment. *Quick Math* (QM) is a practice and assessment system of elementary and middle school mathematics. The system is used to assess and strengthen procedural and representational fluency with concepts of the number system and operations with numbers. Questions in the system are generated on-the-fly from item models, which are templates or schemas of problems that can be instantiated with specific numerical or other parameters to create an actual exercise (Bejar, 1993). Both the accuracy and speed of a response can be used by the system to continually update latent trait estimates of student mastery and fluency of core mathematical concepts and procedures. The system is part of the CBAL™ (Cognitively Based Assessment of, for, and as Learning) research initiative to develop assessments that maximize the potential for positive effects on teaching and learning (Arieli-Attali & Cayton-Hodges, 2014; Bennett, 2011). As a formative assessment, the system can provide immediate feedback on the correctness of responses and overall mastery and fluency levels. Most of the item models used in this study require the student to type a number as an answer, while others require the selection of a response, and some have a hybrid format, where students select objects on the screen to construct their response.

Item bank. The item bank for this study consisted of 400 items generated from 50 item models, with eight instances per model. These items were previously piloted with middle school students

Table 1. Means (and Standard Deviations) of Study Scores.

	No feedback			Feedback		
	F50 (<i>n</i> = 128)	A50 (<i>n</i> = 131)	A70 (<i>n</i> = 130)	F50 (<i>n</i> = 132)	A50 (<i>n</i> = 133)	A70 (<i>n</i> = 135)
% correct	0.50 (0.22)	0.52 (0.12)	0.71 (0.11)	0.56 (0.21)	0.54 (0.13)	0.71 (0.11)
QM	0.24 (1.12)	0.25 (1.21)	0.43 (1.24)	0.52 (0.98)	0.42 (1.29)	0.44 (1.26)
QM-EV	0.25 (1.06)	0.26 (0.97)	0.43 (1.00)	0.51 (0.93)	0.40 (1.03)	0.44 (1.01)
Slowness	0.94 (0.32)	1.06 (0.32)	1.07 (0.38)	0.81 (0.35)	0.94 (0.37)	0.96 (0.36)
Engagement	3.11 (0.75)	2.99 (0.77)	3.18 (0.81)	3.08 (0.75)	2.95 (0.85)	3.39 (0.83)
Anxiety	2.11 (0.63)	2.37 (0.62)	2.02 (0.58)	2.20 (0.60)	2.31 (0.61)	2.01 (0.54)

Note. QM-EV = QM scores with variance set to 1 for each test type. QM = *Quick Math*.

and showed high levels of reliability (Cronbach’s alpha of .96 for 20 min of testing) and validity coefficients, with correlations of .80 with state assessment scores (Attali & Arieli-Attali, 2015). Based on these data, a latent trait model for accuracy of response, similar to a one-parameter item response theory (IRT) model, was developed using a generalized linear mixed model with persons, item models, and instances (nested within models) all defined as random effects (see Janssen, Schepers, & Peres, 2004, for a description of this kind of model). In this study, student ability estimates (labeled henceforth QM scores) were estimated using the expected a posteriori (EAP) method with a standard normal prior and using item instance parameters from the accuracy model.

A separate but similar generalized mixed model was also used to estimate the expected response time for each item based on the data collected in an earlier study (Attali & Arieli-Attali, 2015). To compute time-on-task scores, the deviation of an examinee’s response time from the expected response time for an average examinee estimated in an earlier study (Attali & Arieli-Attali, 2015) was averaged across all items answered by the examinee. For these time-on-task scores, higher values represent slower responses.

Test construction. To construct the FIT QM test for the study, a small pilot was conducted with 10 students from the school who did not participate in the main study. These students completed an adaptive QM assessment (described in more detail below). Using their QM ability estimates together with their state assessment scores and the distribution of state assessment scores of all the students in the school, the expected mean QM score for the school was estimated. A FIT was then constructed with an average item difficulty that was set to be equal to the expected average ability estimate obtained for the school based on the pilot sample. As a consequence, the *expected* percent correct for this fixed test was 50%. The actual average QM score of all the students who took the fixed test was slightly higher than expected, with an average percent correct of 53% (see Table 1). The fixed test was further specified to include 40 items, with 20 item models and two instances per model. The two instances of each model were separated by 10 or more items from other item models. This led to a set of items covering a representative set of content areas and topics with a wide range of difficulty levels. The distribution of the item difficulty IRT parameter ($M = .25$, $SD = .88$) for the 40 items ranged between -1.74 and 1.93 and the middle 50% item difficulty parameters ranged between $-.33$ and $.85$.

The adaptive tests in this study were based on the principle of selecting items that have a specific expected probability for a correct answer, either 50% (regular adaptive test, labeled as A50) or 70% (the easier adaptive test, labeled as A70). This expected probability was derived from the ability estimate of each student during the test and the difficulty parameter of a

particular item. Additional item selection rules specified no more than two instances from each item model, which is similar to the restrictions when constructing the FIT.

Engagement questionnaire. The posttest engagement questionnaire was adapted from the Questionnaire on Current Motivation (QCM; Rheinberg, Vollmeyer, & Burns, 2001; for an English version, see Vollmeyer & Rheinberg, 2006). All questions were presented as 5-point Likert-type scale items (*strongly agree, agree, neutral, disagree, and strongly disagree*). The questions were *I enjoyed today's test, The questions in today's test were interesting, I am eager to know how I performed in today's test, I tried hard to answer today's questions, and It was fun answering today's questions*. Cronbach's alpha internal consistency was .75 for these items. An engagement measure was computed as the average response over these items with a scale of 1 to 5, with 5 representing the highest engagement level.

Anxiety questionnaire. The posttest anxiety questionnaire was adapted from Attali and Powers (2010). The questionnaire includes 12 adjectives (Calm, Tense, Worried, Secure, Frightened, Anxious, At Ease, Nervous, Content, Jumpy, Pleasant, Confused), and respondents are asked to indicate, on a 4-point Likert-type scale (*not at all, a little, moderately, very much*), how well each of these adjectives describe their feelings during the test they just completed. Cronbach's alpha internal consistency was .85 for these items. An anxiety measure was computed as the average response over the 12 adjectives, after reversing the scores of positive ones (Calm, Secure, At ease, Content, Pleasant).

State assessment achievement levels. State assessment (the Florida Comprehensive Assessment Test, FCAT) achievement levels in mathematics, ranging from 1 (lowest) to 5 (highest), were available for all participants. The lowest three levels are defined as "little," "limited," or "partial" success with the content of the state standards. The fourth level is defined as success with the content of the standards, and the fifth level is defined as success with the most challenging content of the standards. These achievement levels were used as an ability estimate independent of QM scores. The QM scores were highly correlated with achievement levels ($r = .81$), evidence for their validity. Because a relatively small number of participants achieved the two highest levels (19%), these two levels were combined into one category (labeled Level 4) for the purpose of analyses.

Design

A 2 (feedback) $\times$ 3 (test type) factorial design was used in the study. Under the feedback condition, examinees were informed of the correctness of each answer immediately after it was submitted, whereas under the no-feedback condition, no feedback was provided. The three test types in this study were the FIT with an expected difficulty of 50% (F50), the adaptive test with an expected difficulty of 50% (A50), and the easier adaptive test with an expected difficulty of 70% (A70). Stratified random assignment of students to one of the six conditions of the study was accomplished by ordering students by the school classroom indicator, math grade, gender, and ethnicity, and sequentially assigning them to one of the six conditions.

Procedures

Participants completed the study in the school's computer lab, which was equipped with Windows PCs, screen resolutions of 1024×768 or higher, and mostly Firefox Version 31 web browsers. Students were asked to log into the online testing system using a unique numeric ID that was preassigned to them. On the first screen, examinees were instructed that they were

going to answer 40 mathematics questions that will involve whole number, fractions, decimals, and percentages. They were given general instructions on how to answer questions (click on options or type numbers). Examinees in the feedback condition were also told that they would receive immediate feedback about the correctness of their response. Examinees were asked to try their best to answer each question correctly as quickly as possible. Examinees were expected to complete the study in one class period (45 min), which was determined based on results of previous data collections with middle school students.

Each test item appeared on the screen one after the other. Examinees typed or selected their answer and clicked on a button labeled "Check Answer" in the feedback condition and "Next Question" in the no-feedback condition. After submitting the answer in the feedback condition, the responses were automatically scored, examinees were presented with the message "Correct!" or "Your answer is incorrect" followed by the correct answer, and the button label changed to "Next Question." Once the test was completed, the anxiety questionnaire was presented followed by the engagement questionnaire. The session ended with a thank-you page notifying students that they have successfully completed the test.

Analyses

First, percent correct scores were examined to verify that average performance corresponded to expectations (50% or 70%) and that the standard deviation of performance was lower for CAT conditions than for the FIT condition. Next, the correlations among the dependent variables in this study, namely engagement, anxiety, time-on-task, and QM scores were examined.

The research questions concerning test engagement, anxiety, time-on-task scores, and QM scores were examined with 2 (feedback type) $\times$ 3 (test type) $\times$ $(4 \text{ state assessment levels})$ ANOVAs on these measures as dependent variables.

The analysis on QM scores with state assessment ability levels was done after transforming the QM scores to create scores had the same standard deviation (set arbitrarily to 1) under each of the test types (but with the means unchanged), to account for possible statistical artifact associated with the variance of QM scores within each test type (see Appendix A for clarifications). Then these transformed QM scores were used as dependent variable in the ANOVA.

Results

The first row of Table 1 shows that the different test types had the intended effect on proportion correct. In terms of average performance, F50 and A50 tests were close to 50%, especially with no feedback, and A70 was close to 70%. In terms of the variability of percent correct scores, the fixed tests show standard deviations which are twice as large as those of the adaptive tests. Nevertheless, the adaptive test conditions did not completely eliminate the variability in percent correct scores. As a consequence, higher ability examinees tended to answer more questions correctly even in adaptive testing conditions, although to a lesser degree than in the FIT.

Table 2 shows that the four study scores were weakly correlated. Students with higher QM scores tended to spend relatively less time on task ($r = -.32$), and students reporting more anxiety tended to have lower QM scores ($r = -.20$) and report lower engagement ($r = -.32$).

Effects on Engagement, Anxiety, Time-on-Task, and QM Scores

For each of the four measures, a 2 (feedback type) $\times$ 3 (test type) $\times$ $(4 \text{ state assessment levels})$ ANOVA was performed.

Table 2. Correlations Between Study Scores.

	Slowness	Engagement	Anxiety
QM	-.32*	-.09*	-.20*
Slowness		.05	.14*
Engagement			-.32*
Anxiety			

Note. QM = Quick Math.

* $p < .05$.

For engagement scores, a significant test type main effect was found, $F(2, 765) = 12.29, p < .01, \eta_p^2 = .031$, and a post hoc Tukey's test showed that A70 engagement ($M = 3.30, SE = .05$) was significantly higher than either A50 ($M = 2.97, SE = .05$) or F50 ($M = 3.10, SE = .05$). The ability effect was also significant, $F(3, 765) = 6.86, p < .01, \eta_p^2 = .025$, with lower engagement for higher ability Levels 3 and 4 (see Figure 1). The feedback main effect, as well as any interaction between main effects, was not significant.

For anxiety scores, a test type main effect was found, $F(2, 765) = 20.57, p < .01, \eta_p^2 = .051$, and a post hoc Tukey's test showed that A70 anxiety ($M = 2.01, SE = .04$) was significantly lower than F50 ($M = 2.16, SE = .04$), which was in turn significantly lower than A50 ($M = 2.35, SE = .04$). The ability effect was also significant, $F(1, 765) = 4.54, p < .01, \eta_p^2 = .018$, with lower anxiety for higher ability Level 4 (see Figure 1). The feedback main effect, as well as any interaction between main effects, was not significant.

For time-on-task scores, a more complex picture emerged. First, all three main effects were found to be significant. A test type main effect was found, $F(2, 765) = 18.06, p < .01, \eta_p^2 = .045$, with lower F50 time-on-task scores ($M = .88, SE = .02$) than either A50 ($M = 1.01, SE = .02$) or A70 ($M = 1.04, SE = .02$). In other words, examinees tend to spend less time answering items in the F50 condition than in either A50 or A70. A feedback main effect was found, $F(1, 765) = 26.32, p < .01, \eta_p^2 = .033$, with lower time-on-task with feedback ($M = .91, SE = .02$) than that without feedback ($M = 1.03, SE = .02$). This is consistent with previous results and was explained (Attali & Powers, 2010) through the psychological effect that expecting immediate feedback has on the ease of submitting an answer, especially for constructed-response (CR) items. An ability main effect was found, $F(3, 765) = 41.49, p < .01, \eta_p^2 = .140$, with lower time-on-task for higher ability levels. However, the Test Type $\times$ Ability interaction was also significant, $F(6, 765) = 4.59, p < .01, \eta_p^2 = .035$, showing that the lower time-on-task of F50 (compared with either A50 or A70) is due only to a significant difference at the lowest ability level 1 (see Figure 1). Furthermore, the three-way interaction (Test Type $\times$ Ability $\times$ Feedback) was also significant, $F(6, 765) = 2.79, p < .01, \eta_p^2 = .021$, showing that the lower time-on-task of F50 in the lowest Level 1 (compared with either A50 or A70) is due only to a significant difference for the no-feedback condition (see Figure 2). In other words, a test type effect, with lower time-on-task (i.e., less time on task) for F50 than for A50 or A70, was found only for the lowest ability level and for the no-feedback condition.

For transformed QM scores (with equal standard deviations for each of the test types), a feedback main effect was found, $F(1, 765) = 10.07, p < .01, \eta_p^2 = .013$, with higher QM scores in the feedback condition ($M = .39, SE = .03$) than in the no-feedback condition ($M = .25, SE = .03$). The ability effect was also significant, $F(3, 765) = 450.92, p < .01, \eta_p^2 = .639$, with higher QM scores for higher ability (see Figure 1). The test type main effect, as well as any interaction between main effects, was not significant. Nevertheless, a post hoc Tukey's test on the

Figure 1. Adjusted means (and SE) across achievement levels and test types.

Feedback $\times$ Test Type interaction, $F(2, 765) = 1.75, p = .17, \eta_p^2 = .005$, revealed a significant increase in QM scores for F50 from no-feedback to feedback condition (see Figure 3).

Discussion

The purpose of this study was to explore possible effects of test adaptivity and difficulty of the adaptive test on several measures of test motivation and test performance. The A50 is compared with the F50 to examine effects of adaptivity, and is compared A70 to examine effects of an easier CAT. The math ability tests were administered either without feedback (standard practice) or with immediate feedback of the correct answer to ensure that examinees would receive accurate feedback about their performance throughout the test.

In summary, results indicated almost no support for an advantage of adaptivity (A50 vs. F50) and some support for an advantage of an easier CAT over standard CAT (A70 vs. A50). In the comparison between CAT and FIT (A50 vs. F50), the results were mixed. No differences in test engagement were found, but F50 showed lower (better) test anxiety. On the contrary, time on

Figure 2. Adjusted means (and SE) for slowness across achievement levels, test types, and feedback conditions.

task (a measure of persistence) was lower (or worse) under F50 for lower ability examinees when no feedback was provided. In addition, performance (QM test scores) for F50 was lower without feedback than with feedback (but not lower than A50). In the comparison between a harder and easier CAT (A50 and A70), an advantage for A70 was found in terms of test engagement, but no differences were found in terms of anxiety, time on task, or test performance.

The role of feedback was surprising. It was introduced as a way to inform examinees about their performance in the adaptive test, but had no effect on the two CAT variants. Instead, it had a minor beneficial effect on the FIT, increasing time on task for lower ability examinees and test performance for all examinees. In addition, feedback had a beneficial main effect on test performance, across test types and independent of examinees' prior math achievement levels.

Examinee ability (as measured by state assessment achievement levels) had an overall effect on each of the four measures. Higher ability examinees tended to report less anxiety in response to the test, but also reported less engagement and spent less time on task. They also performed better on the test. As mentioned earlier, the only differential test type effect that ability showed was the shorter time on task for lower ability examinees in the F50 test.

The results on test performance are generally consistent with previous research where no significant test type effect was found. However, some differences between the study's results and previous research on measures of motivation and anxiety should be noted. Some previous studies reported higher motivation under CAT for low-ability students (Betz, 1977; Betz & Weiss, 1976) and others report higher motivation under CAT in general (Arvey et al., 1990; Pine et al., 1979), but differences in test engagement between A50 and F50 are not found. However, Betz

Figure 3. Adjusted means (and SE) for QM scores across test types and feedback conditions.
Note. QM = Quick Math.

(Betz, 1977; Betz & Weiss, 1976) found higher anxiety on the CAT, consistent with the findings of higher anxiety for A50 over F50.

The current study was motivated by two design decisions that were hypothesized to potentially mediate test type effects. First, this study used a mathematical ability assessment, whose tasks are more cognitively demanding than tests of declarative knowledge (knowledge of medical technology and vocabulary) used in earlier research (e.g., Betz & Weiss, 1976). It was speculated that with these tasks, possible engagement effects (and as a consequence, possible performance effects) would be stronger on the mathematical ability assessment because they require more effort to answer. In this respect, the relative lack of test type effects on the mathematical ability assessment can be interpreted as stronger evidence for the equivalence of the different test types in terms of test performance and test motivation.

The second design decision was the manipulation of immediate feedback. Immediate feedback about the correct answer (absent from most tests, including CATs) seems particularly important for the adaptive test design that seeks to control item difficulty, because feedback provides unambiguous information to examinees about their success throughout the test and therefore, with immediate feedback, examinees would know whether they are answering most items correctly or incorrectly. This is in contrast to the ambiguous information provided in a CAT without feedback, where examinees can try and estimate the difficulty of items to gauge whether they answered the previous items correctly. Nevertheless, the presence or absence of feedback did not seem to affect test type effects. In particular, test type effects were not stronger under feedback conditions. At the same time, feedback had a beneficial (main) effect on performance. But similar to the point made about test content, the lack of test type effects under feedback conditions can be interpreted as stronger evidence for the equivalence of the different test types.

In a similar vein, it was noted that the sample sizes in this study were larger than previous studies on this topic, and therefore the relative lack of test type effects can again be interpreted as stronger evidence for the equivalence of the different test types.

How can the consistent lack of beneficial engagement and performance effects for CAT over FIT be explained, despite the differences in testing protocols and patterns of test difficulty that could affect many examinees? One possibility that was explored in the past (Ortner & Caspers, 2011; Ortner, Weißkopf, & Gerstenberg, 2013; Ortner et al., 2014; Tonidandel & Quiñones, 2000; Tonidandel, Quiñones, & Adams, 2002) is that, in fact, certain features of a CAT could result in negative reactions by examinees. One such feature is the inability to skip items in a CAT because items are assigned to examinees based on responses to previous items. Ortner and Caspers (2011) also hypothesized that tests containing mainly items with medium relative difficulty could have negative effects on test performance for examinees high in test anxiety. The possibility that certain groups of examinees would be adversely affected by a CAT raises the potential explanation that the lack of overall test type effects result from contradictory negative and positive effects by different groups. Unfortunately, previous studies and the results of the current study do not support this possibility—no study found contradictory test type effects by different groups (based on ability or test anxiety).

Are Test Experiences Under CAT Really Different?

As an alternative explanation for this phenomenon, the authors would like to raise the possibility that the test experiences of examinees under the CAT and FIT test types may not be that different. To explain this point, the authors would like to draw attention to the variability in the number of correct answers across examinees under CAT and FIT. In the no-feedback condition, the *SD* of percent correct scores in the CAT (A50) was .12, compared to an *SD* of .22 for the percent correct scores in the FIT (see Table 1). Although the FIT *SD* was reduced by about 50%, it is still considerable and suggests that under the CAT condition examinees with low QM scores still answered less items correctly than examinees with high QM scores.

One possibility for this result is that the CAT implementation was not entirely successful in assigning the most appropriate items to examinees due to limitations in the item bank. The percentage of correct answers for an examinee taking a CAT is certainly influenced by the quality of the item bank. With a small item bank, it may be difficult to assign consistently easy items to a low-ability examinee and difficult items to a high-ability examinee. Similarly, with items that are not discriminating well, it will take time (or items) to “find” the true ability of an examinee. However, even with a large item bank composed of high-quality items, it takes time (or items) to adjust the estimate of ability, and the more extreme the ability is, the more time it will take. As tests have typically just a few dozen items, it is difficult to avoid some variability in percent correct scores that is related to examinee true ability.

To illustrate this point, a simulation was performed with a “perfect” CAT, one with an infinite item bank where item difficulty is always perfectly matched to current ability estimate (see Appendix B). As expected from a highly reliable test, the correlation between true scores and final ability estimates was very high (.95). However, a considerable variability in percent correct scores remained (only slightly lower than the study’s empirical results), and the correlation between final ability estimates and the number of correct answers was very high (.81). This simulation demonstrates that even under optimal CAT conditions, high-ability examinees still answer most items correctly, low-ability examinees still answer most items incorrectly, and the nominal success of examinees (how many items they answered correctly) is still highly correlated with their ability.

As a consequence, the test experience of examinees in a CAT may not be very different from the experience in a FIT. This is especially true for examinees who do not possess extreme ability levels, which by definition constitutes the vast majority of examinees. Note, however, that this conclusion is most readily applicable to typical FITs that are constructed with a range of

item difficulties (as was the case in this study). In principle, a FIT could be constructed with no variability in item difficulty—for example, only middle-difficulty items or even only extreme-difficulty items. This would result in greater discrepancies in experience between the CAT and FIT for a greater proportion of examinees. This, however, is not practical from a psychometric approach, as it will further deteriorate the measurement accuracy for examinees with lower or higher ability levels. This, in turn, will lead to a lower reliability coefficient than a FIT with items that are more varied in terms of item difficulty.

The other two manipulations that were explored in this study are easier to detect by examinees and seemed to have a larger impact on examinees' reactions to the test. The first, immediate feedback, had a beneficial effect on test performance, but not on engagement or anxiety. The second, easiness of the adaptive test, was beneficial for test engagement and anxiety, but not for performance.

These results could have implications for the design of adaptive learning systems. These systems both guide students through a curriculum of instructional activities, and monitor step-by-step progress on an activity (Van Lehn, 2006). As a consequence, adaptive assessments are essential components of these systems. However, despite a growing interest in these types of assessments for learning, the authors still lack strong evidence of their effectiveness in promoting student achievement (Bennett, 2011). Moreover, even a basic component of formative assessment, the provision of feedback to learners, is not well understood (Shute, 2008) and the results of this study suggest that immediate feedback has a positive effect on students' performance regardless of their ability levels.

In conclusion, the authors propose that test adaptivity, although providing substantial psychometric advantages, may not in fact result in substantially different test experiences for most examinees. This would explain the lack of test type effects in this study, as well as in previous research. This intriguing possibility could be explored in future research. For example, it would be interesting to ascertain the degree to which examinees can identify the unusual item selection procedures when administered a CAT without being told of these procedures in advance. To the extent that examinees indeed do not experience a CAT differently than a FIT, the appeal of adaptive testing as a more efficient alternative to traditional fixed-item testing should increase.

Appendix A

Adjustment of Possible Statistical Artifacts Associated With the Variance of Quick Math (QM) Scores Within Each Test Type

The analysis on QM scores with state assessment ability levels is likely affected by a statistical artifact associated with the variance of test scores based on each test type. First, note that less reliable measures are expected to have lower score variance. Therefore, F50 is expected to have lower score variance than A50 because of the lower measurement efficiency in fixed-item tests (FIT) than computer adaptive test (CAT). In other words, F50 scores are more regressed than A50 scores. Second, because ability as measured by the state assessment ability levels is highly correlated with QM scores, the low variance F50 scores will be more regressed on ability than the high variance A50 scores. As a consequence, mean F50 scores for *low* ability level examinees are expected to be *higher* than those for A50 scores, due to the more regressed nature of F50 scores. Conversely, mean F50 scores for *high* ability level examinees are expected to be *lower* than those for A50 scores, due to the same regressed nature of F50 scores. These two layers of issues are likely to result in an artificial interaction effects between test type and ability levels that may not actually exist.

Appendix B

Variability of Percent Correct in a Simulation of a Perfect CAT

To illustrate the fact that, even under optimal circumstances, some variability in percent correct scores will remain under a CAT, a simulation was performed. In it, a “perfect” CAT was administered, one with an infinite item bank where item difficulty is always perfectly matched to current ability estimate. Specifically, the true score of 1,000 simulated examinees was first systematically drawn from $N(\mu = 0.25, \sigma = 1.25)$, to have the same parameters as the A50 sample. Then each simulee was administered a perfect CAT with 40 items as described above, and correctness of responses was randomly generated based on relative item difficulty (relative to the true score). The correlation between true scores and final ability estimates was .95, indicating that the simulated CAT was highly reliable. However, the correlation between final ability estimates and the number of correct answers (these could range from 0 to 40) was .81, an extremely high value considering the common belief that a CAT is supposed to control the number of items answered correctly across all ability levels. Moreover, similar to the empirical results, a considerable variability in percent correct scores was observed for the perfect simulated CAT. With an $SD = .10$, it was only slightly lower than the empirical results ($SD = .12$).

Acknowledgment

The authors would like to thank the participating school and coordinators for helping with the data collection. This article also benefited greatly from valuable comment from Dr. Donald Powers, Dr. Diego Zapata, Dr. Joseph Rios, and Dr. Lydia Ou Liu from Educational Testing Service (ETS), Dr. Hua-Hua Chang from University of Illinois at Urbana Champaign, as well as two anonymous reviewers. All the views and opinions presented in this article were solely those of the authors, but do not reflect those of ETS.

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: This study was funded by ETS Internal Research Program.

References

- Arieli-Attali, M., & Cayton-Hodges, G. A. (2014). *Expanding the CBAL competency model for mathematics assessments and developing a Rational Number learning progression* (Research Report 14-08). Princeton, NJ: Educational Testing Service.
- Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*, 695-716.
- Atkinson, J. W. (1957). Motivational determinants of risk taking behavior. *Psychological Review, 64*, 359-372.
- Attali, Y., & Arieli-Attali, M. (2015). Gamification in assessment: Do points affect test performance? *Computers & Education, 83*, 57-63.
- Attali, Y., & Powers, D. (2010). Immediate feedback and opportunity to revise answers to open-ended questions. *Educational and Psychological Measurement, 70*, 22-35.

- Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323-359). Hillsdale, NJ: Lawrence Erlbaum.
- Bennett, R. (2011). *CBAL: Results from piloting innovative K-12 assessments* (Research Report 11-23). Princeton, NJ: Educational Testing Service.
- Bergstrom, B. (1992, April). *Ability measure equivalence of computer adaptive and paper and pencil tests: A research synthesis*. Paper presented at the Annual Meeting of the American Education Research Association, San Francisco, CA.
- 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). Minneapolis: Department of Psychology, University of Minnesota.
- Chang, H. H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222.
- Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 62, 369-383.
- DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, 13, 55-77.
- Häusler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science Quarterly*, 50, 75-87.
- Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 35, 69-81.
- Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 189-212). New York, NY: Springer.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119, 254-284.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Lawrence Erlbaum.
- Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive testing conditions. *Journal of Educational Measurement*, 31, 251-263.
- Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34, 207-218.
- Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114, 449-458.
- Ortner, T. M., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*, 27, 157-163.
- Ortner, T. M., Weißkopf, E., & Gerstenberg, F. X. (2013). Skilled but unaware of it: CAT undermines a test taker's metacognitive competence. *European Journal of Psychology of Education*, 28, 37-51.
- 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, 48-56.
- Penk, C., & Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences*, 42, 27-35. doi:<http://dx.doi.org/10.1016/j.lindif.2015.08.002>
- Pine, S. M., Church, A. T., Gialluca, K. A., & Weiss, D. J. (1979). *Effects of computerized adaptive testing on black and white students* (Research Report 79-2). Minneapolis: Department of Psychology, University of Minnesota.

- Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, 95(4), 667-686.
- Ponsoda, V., Olea, J., Rodriguez, M. S., & Revuelta, J. (1999). The effect of test difficulty manipulation in computerized adaptive testing and self-adapted testing. *Applied Measurement in Education*, 12, 167-184.
- 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, 249-273.
- Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen [QCM: A questionnaire to assess current motivation in learning situations]. *Diagnostica*, 47, 57-66.
- Schunk, D. H., Meece, J. R., & Pintrich, P. R. (2014). *Motivation in education: Theory, research, and applications* (4th ed.). Boston, MA: Pearson.
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78, 153-189.
- Tonidandel, S., & Quinones, M. A. (2000). Psychological reactions to adaptive testing. *International Journal of Selection and Assessment*, 8, 7-15.
- 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, 320-332.
- Van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. St. Paul, MN: Assessment Systems Corp.
- Van Lehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16, 227-265.
- Vollmeyer, R., & Rheinberg, F. (2006). Motivational effects on self-regulated learning with different tasks. *Educational Psychology Review*, 18, 239-253.
- Wainer, H. (2000). Introduction and history. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, & R. J. Mislevy (Eds.), *Computerized adaptive testing: A primer* (pp. 1-21). Hillsdale, NJ: Lawrence Erlbaum.
- Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2007). A meta-analysis of testing mode effects in Grade K-12 mathematics tests. *Educational and Psychological Measurement*, 67, 219-238.
- 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). Minneapolis: Department of Psychology, University of Minnesota.
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21, 361-375.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68-81.
- Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, 2, 1-17.
- Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, 15, 27-41.
- Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education*, 8, 341-351.