Computers & Education 100 (2016) 45–55

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

The effects of computer self-efficacy, training satisfaction and test anxiety on attitude and performance in computerized adaptive testing

Hong Lu [a, *], Yi-ping Hu [a], Jia-jia Gao [a], Kinshuk [b]

[a] College of Communication, Shandong Normal University, China
[b] Faculty of Science and Technology, Athabasca University, Canada

ARTICLE INFO

Article history:
Received 20 October 2015
Received in revised form 28 April 2016
Accepted 29 April 2016
Available online 30 April 2016

Keywords:
Computer self-efficacy
Training satisfaction
Test anxiety
Computerized adaptive testing
Structural equation model

ABSTRACT

This study focused on test-takers' psychological effects on computerized adaptive testing (CAT). The development and implementation of CAT were based on item response theory (IRT), and two-parameter logistic model was chosen for the items. The total of 268 students from a high school in Jinan took part in the English adaptive test. A structural equation model was used to examine the potential connections among a series of individual variables (computer self-efficacy, training satisfaction, test anxiety, CAT attitude and CAT performance). The findings revealed significant positive paths from computer self-efficacy and training satisfaction to CAT attitude, as well as a negative path from test anxiety to CAT performance. Furthermore, there was significant correlation between the residual variances of CAT attitude and CAT performance. Thus, it could be seen CAT might produce an unfair disadvantage for test-takers with higher test anxiety. The relevant research and implications were further discussed.

1. Introduction

As information technology has become increasingly more prevalent and accessible for use in student assessment, innovative test delivery models are adopted to collect, analyze, and report student-level data. Among these models, computerized adaptive testing (CAT) based on item response theory (IRT) has been attracting more and more attention. The basic idea of CAT is that test items are selected by the computer to individually match the ability level of each student (Wainer, 2000). In this manner, the test is tailored to each student. There are some benefits associated with CAT, and it is logical to see why testing experts are making a push toward this testing modality. For example, by using more precise and efficient assessments that take less time to complete, teachers and students will get test results that are either just as accurate as traditional tests or more accurate. In addition to this, the tests tailor each question to the knowledge and abilities of the students, thereby theoretically keeping them appropriately challenged and more likely to stay engaged (Wainer, 2000). Based on the above mentioned advantages, CAT is becoming more and more common in high-stake assessment. For instance, the Graduate Record Examination (GRE), the nursing licensing exam, and the Graduate Management Admission Test (GMAT), are all now primarily offered in CAT. Additionally, in the U.S., many states were moving to put in place online testing tied to the common

* Corresponding author.
E-mail address: sdnulh@163.com (H. Lu).

http://dx.doi.org/10.1016/j.compedu.2016.04.012
0360-1315/

core state standards in 2014–15, at least 20 states among them indicated they would plan to use new computer-adaptive versions of the tests (Davis, 2012). Moreover, the Smarter Balanced Assessment Consortium has received federal funding to develop English/language arts and mathematics adaptive tests for the common standards. He said his assessment would feature high-tech, interactive questions that incorporated video and graphics, and were designed both to identify what students knew and to be more engaging (Davis, 2012).

1.1. Review of literature

The trend that CAT was an evolutionary step toward future testing methodologies resulted in a growing number of studies dedicating to it. Most of them investigated the technical aspects of CAT, such as comparison of different item-selection methods (Finkelman, Kim, Weissman, & Cook, 2014; He, Diao, & Hauser, 2014; Wang, 2013a; Yao, 2012), item pool construction (He & Reckase, 2014; Lee & Dodd, 2012), test stopping rules (Choi, Grady, & Dodd, 2011; Wang, Chang, & Boughton, 2013; Yao, 2013). However, only a few studies dealt with CAT's psychological effects on test-takers. In early studies, chief among them was that it might increase the student's interest and motivation for taking the test. For instance, Weiss and Betz (1973) indicated that adaptive testing was suggested to avoid boredom for test-takers with high ability and prevent test-takers with low ability from experiencing anxiety. Johnson and Mihal (1973) found that blacks performed better on adaptive testing. Weiss (1975) found similar motivational effects when feedback on the correctness of a response was provided. These results seemed to suggest that in some cases CAT might be more motivating or less anxiety-producing than conventional testing.

In recent years, undesirable psychological reactions to CAT were discussed as following: Tonidandel and Quinones (2000) explored how specific aspects of adaptive testing influence test-takers' reactions. Fifty-three undergraduates were presented with descriptions of hypothetical selection tests manipulated to reflect characteristics of adaptive tests that differed from traditional paper-and-pencil tests (P&P). The results demonstrated that certain features of adaptive tests, such as the inability to skip questions, review items, or go back and change answers, might adversely impact test-takers' psychological reactions. Ortner and Caspers (2011) investigated the effects of test anxiety on test performance using computerized adaptive testing versus conventional fixed item testing. A total of 110 students from a German secondary modern school were tested. Findings showed that, when confronted with an adaptive matrices test, test-takers with high test anxiety had lower test scores compared to persons with low test anxiety. That was to say, adaptive testing might lead to a bias that produced a disadvantage for test-takers with higher test anxiety. In another study, Ortner, Weiβkopf, and Koch (2014) examined the effects of computerized adaptive testing versus computerized fixed item testing of reasoning ability on current motivation. A group of 174 students from two German secondary schools was presented either an adaptive or a fixed version of a matrices test. Less motivation was reported using adaptive testing compared to fixed item testing.

1.2. The present study

The researches published on CAT and psychological effects have yielded mixed results, raising the question of whether or not CAT does support fair assessment procedures for test-takers. For example, some previous researches showed that CAT might be revealed to be unfair with reference to its potential to evoke success-related estimations in high performers, and then the perceived unfairness of CAT had a negative impact on their CAT performance (Ortner et al., 2014). The result contradicted early assumptions generally supported higher fairness for CAT that every test-taker would solve about 50% of the given items correctly independent of ability. The mixed results also made researchers turn their attention to whether CAT was fair or not (Fritts & Marszalek, 2010; Ortner & Caspers, 2011; Ortner et al., 2014; Tonidandel & Quinones, 2000). So the goal of the present study was to further investigate the influence of some individual characteristics on CAT, and provided empirical evidence for fairness or unfairness of CAT.

In order to achieve the goal of this study, three individual characteristics which might have relationship with individual CAT performance or CAT attitude were chosen. The computer was the essential tool during the process of CAT. Therefore, computer self-efficacy should play an important part in applying CAT. In addition, due to the significant difference between CAT and P&P, CAT training was particularly necessary for test-takers. Certainly, training satisfaction was considered to have significant influence on the implementation of CAT. What's more, test anxiety was an essential variable widely studied in the context of various academic achievements (Chapell, et al., 2005; Farooqi, Ghani, & Spielberger, 2012). Thus, computer self-efficacy, training satisfaction, test anxiety, CAT attitude and CAT performance were included in this study to set up a causal model of CAT. The relationships between these latent variables were analyzed by using high-level analysis software as well.

1.3. Research model

Based on previous studies, hypotheses developed to test the effect of the variables of computer self-efficacy, training satisfaction, test anxiety, CAT attitude and CAT performance on each other and their relation to each other were presented below.

1.3.1. CAT attitude and CAT performance

Attitude has been defined as "the favorable or unfavorable response to things, people, places, events or ideas" (Papanastasiou & Zembylas, 2002). McGuire (1985) noted that attitude generally comprised the following three components: cognition (knowledge about an object), affect (feeling about an object), and behavior (tendency to act with or react to an object). Because these three components could be treated independently, here attitude was viewed as cognitive and affective components that form the basis for evaluative judgments. Accordingly, CAT attitude was defined as the feelings that test-takers had towards CAT, which were based on their beliefs about CAT. Of course, CAT attitude was a key impact factor of CAT performance because previous studies had shown that better attitude usually led to a positive commitment (Shen, Wu, & Lee, 2014). Thus, Hypothesis 1 was as following.

H1. CAT attitude had a positive effect on CAT performance.

1.3.2. Computer self-efficacy and CAT attitude

Self-efficacy has been distinguished by Bandura (1993) as a component of students' personal factors related to behavioral changes that often affected students' motivation. It is a self-perception of ability to accomplish an activity. On this basis, Compeau and Higgins (1995) defined computer self-efficacy as "an individual's perceptions of his or her ability to use computers in the accomplishment of a task" (p. 191). It has been a strong predictor of a variety of computing attitudes and beliefs (Celik & Yesilyurt, 2013; Pellas, 2014). The definition could lead to the following hypothesis.

H2. Computer self-efficacy had a positive effect on CAT attitude.

1.3.3. Training satisfaction and CAT performance

Generally speaking, P&P required that all test-takers responded to all same items in the test. Unlike P&P, CAT involved the dynamic selection of items to match the performance of a test-taker during test administration. That was, CAT provided different test item sets for each test-taker based on his/her estimated ability level. Informing about the CAT technique would reduce test-takers' negative effects resulting from divided attention on test-irrelevant cognitions, and test-takers would have more resources at hand for task performance (Ortner & Caspers, 2011). So, the following hypothesis was investigated.

H3. Training satisfaction had a positive effect on CAT performance.

1.3.4. Test anxiety and CAT performance

Test anxiety was defined as "a set of phenomenological, physiological and behavioral responses that accompany concern about possible negative consequences of failure on an exam or similar evaluative situation" (Zeidner, 1998). Spielberger and Vagg (1995) stated that individuals with test anxiety were more prone to react with excessive anxiety (e.g., worry, emotional and physiological arousal) during evaluative situations, such as exams. Numerous studies revealed a negative correlation between test anxiety and performance (Chapell et al., 2005; Iroegbu, 2013; Rezazadeh & Tavakoli, 2009; Trifoni & Shahini, 2011). Therefore, this study constructed Hypothesis 4.

H4. Test anxiety had a negative effect on CAT performance.

1.3.5. Computer self-efficacy, training satisfaction and test anxiety

Most studies investigating the relationship between self-efficacy and anxiety came to the consistent conclusions that they had negative correlation (Paul, Hauser, & Bradley, 2007; Singh, Bhadauria, Jain, & Gurung, 2013; Yukselturk & Bulut, 2007). Moreover, researches on self-efficacy and anxiety also showed that it's easier for individuals with higher self-efficacy to produce feeling of satisfaction (Johnson, Gueutal, & Falbe, 2009; Jung, 2014). On the contrary, it's more difficult for individuals with higher anxiety to produce such feeling (Bolliger & Halupa, 2012; Kim, Han, Lee, Min, Park, & Lee, 2013). So, the following hypotheses were constructed.

H5. Computer self-efficacy had negative correlation with test anxiety.

H6. Computer self-efficacy had positive correlation with training satisfaction.

H7. Test anxiety had negative correlation with training satisfaction.

Base on the above-mentioned hypotheses, the original research model (M1) of the present study was shown in Fig. 1.

2. Methods

2.1. Participants and procedure

The procedure of the present study was divided into three phases. The first phase was to develop an English adaptive testing system. For the purpose of constructing item bank of adaptive testing system, eight high school English teachers developed 500 multiple-choice items concerning knowledge of English listening, grammar, and vocabulary. 420 items were

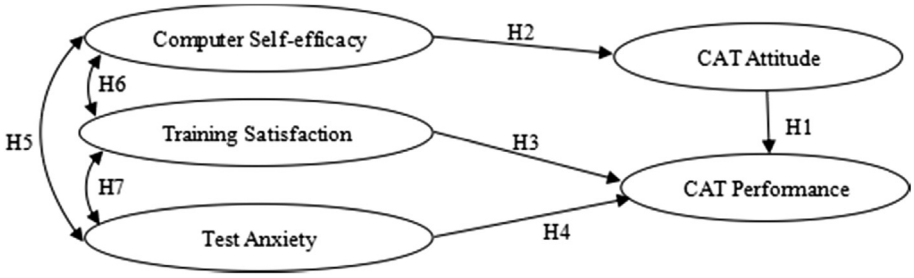

Fig. 1. Path diagram of M1.

drawn to assemble five parallel English tests. Each test consisted of 20 anchor items and 80 independent items. During the process of calibrating item parameters based on IRT, 5672 students in grade 11th from five Chinese high schools in the city of Jinan were recruited to take the pretest. In each high school, students were randomly divided into 5 groups and were assigned one test each. Two-parameter logistic model, which was one of the most popular unidimensional IRT models, was adopted in this study to calibrate the item parameters. And its formula was $P_j(\theta) = e^{Da_j(\theta-b_j)}/[1+e^{Da_j(\theta-b_j)}]$, where j was the sequence number of the test item starting at one, $P_j(\theta)$ denoted the probability that examinee at the ability level θ answered item j correctly, a_j was the discrimination of the jth item, b_j was the difficulty of the jth item, and D was a constant 1.702. Based on the students' responses, item parameters for each of the tests were estimated by applying the Bayesian Expected A Posterior method using BILOGMG 3.0 to model the test response matrix (5672×420). The estimation procedure was configured for a maximum 20 EM cycles. The results turned out that 32 items failed to fit the two-parameter logistic model, and were eliminated according to the value of chi-square and freedom. The parameters of the remaining 388 items were linked on the same scale by the mean and sigma method. Ultimately, 388 items which varied in terms of their discrimination and difficulty constructed item bank of English adaptive testing system. Furthermore, scale unidimensionality was assessed by fitting a one-factor model to the items within each test using SPSS. The results supported the unidimensionality of tests and provided evidence of their appropriateness for IRT modeling.

The second phase was CAT training for 282 students who would take posttest in grade 11th from one high school in Jinan, which was essential because of the unpopularity of the test mode in China. Students were divided into 6 groups to receive the training. It was separated into two stages, and each took about 1.5 h. In the first stage, the students received information related to the basis of CAT, during which the complicated mathematical principles incorporated into IRT were avoided. Firstly, a 20-min introductive video of CAT was presented, and then followed with a 70-min lecture centered on the basic principle of how CAT works. For example, CAT usually begins with an item of medium difficulty. Then it applies a dynamic process of item presentation, and the difficulty of subsequent item is adapted to the test-takers' estimated ability. After two days, the second stage started. At first, the students were informed with the matters needing attention in CAT. For instance, CAT didn't allow test-takers to skip items, review items, or go back to change answers, and test-takers were forced to respond to a question before moving on to the next question. Thirty minutes later, the students were allowed to use an online CAT system for practice, and any question about CAT would be answered by the instructors in this process. In the end of training, 282 students filled out the questionnaire of test anxiety, computer self-efficacy, training satisfaction, and CAT attitude.

The third phase was CAT implementation. At last, 268 students who received CAT training in grade 11th from the same high school in the second phase took part in the posttest (14 students missed the posttest for different reasons). CAT was conducted on notebook computers in a quiet experimental laboratory at the students' school. Approximately 15–20 students were tested simultaneously. Experimental instructions were presented via the notebook computer. The instructions explained to students that they were about to take a fixed-length and time restricted adaptive test which designed to assess their knowledge of English listening, grammar, and vocabulary, and that the test consisted of 36 multiple-choice items (12 items for listening, 12 items for grammar, and 12 items for vocabulary). After the instructions were presented, students had 45 min to take the adaptive test. Additionally, students were told that on completion of the test, their scores would be shared with others. When the process finished, students then received feedback regarding their actual level of performance on the adaptive test.

2.2. Measurement development

2.2.1. Computer self-efficacy

In developing a new measure of computer self-efficacy, reference was made to the work of Compeau and Higgins (1995). There were 10 items in the scale developed by Compeau and Higgins (1995) which represented the potential to use the software in the accomplishment of a task, rather than reflecting simple component skills. However, with the development of information technology, some changes should be done to refine the original scale. For example, people could get help from the Internet instead of consulting software manuals or asking someone around. Five experts from the field of computer,

educational psychology, and educational technology were asked to consider the content of the item, the level of difficulty, and overall comprehensiveness of the original scale. Results of the expert review were then used to revise the original items and shorten the form to 3 items (see Appendix). Each of the items was still preceded by the phrase "I could complete the job using the software package." A five-point Likert-type response was employed and respondents were asked to indicate the degree to which they felt not at all confident (1) to totally confident (5). All items were positively-worded statements and high scores indicated a high degree of confidence to use computers.

2.2.2. Training satisfaction

In order to obtain a short form inventory of training satisfaction, the present study simplified the training satisfaction rating scale which was developed by Tello, Moscoso, Garcia, and Chaves (2006). The original scale contained 12 items grouped into three dimensions: (1) objectives and content, (2) method and training context, and (3) usefulness and overall rating. A validity study was carried out using expert judge which was constituted by six experts from different universities and private training firms. The experts evaluated each item with respect to its representativeness and utility. As a result, a final 3 items, five-point (1 = "totally disagree", 5 = "totally agree") inventory (see Appendix) was selected whose content was representative and useful to measure original three dimensions according to the content validity study.

2.2.3. Test anxiety

Items for measuring test anxiety (see Appendix) were adapted from Taylor and Deane's (2002) study. In their study, the short form test anxiety inventory (TAI) which consisted of 5 four-point items was derived from a full form original TAI (Spielberger et al., 1980). Each item had the following response scale: 1 = "almost never," 2 = "sometimes," 3 = "often," 4 = "almost always." Internal consistency and concurrent and construct validity were assessed in hypothetical and actual examination conditions. The short form TAI produced optimal reliability and validity, and a balance of items from the worry and emotionality subscales of the original TAI.

2.2.4. CAT attitude

Twelve graduate education technology students generated statements in terms of cognitive and affective components about CAT, leading to 6 items for inclusion in the original scale of CAT attitude. Subsequently, forty-eight graduate students from educational research method classes used a five-point Likert scale to indicate their level of agreement or disagreement with each of these items. Items retained for the final scale were those that: (a) best discriminated between the 25 percent of the subjects with the highest and lowest total scores, and (b) demonstrated the highest item-total correlations. The final version of CAT attitude (see Appendix) consisted of 3 five-point Likert items which expressed positive attitudes towards CAT. Possible scores on the final scale of CAT attitude ranged from a minimum score of 3 (indicating an extremely negative attitude toward CAT) to a maximum score of 15 (indicating an extremely positive attitude toward CAT).

2.2.5. CAT performance

According to IRT, CAT performance was not based on the number of items answered correctly, but rather, on which items were answered correctly. In other words, CAT performance would be greater if he or she correctly answered difficult questions as opposed to easy questions. In this study, CAT performance eventually was presented on the basis of percentile rank of performance on the test.

2.3. Statistical analysis

The research model shown in Fig. 1 was analyzed primarily using structural equation model (SEM), supported by LISREL 8.70 software. Data analysis was carried out in accordance with a two-stage methodology: the measurement model and the structure model. The first step was to assess the reliability and construct validity for the five measurement elements. In the second step, the paths between the latent constructs were modified with the structural equation model.

3. Results

3.1. Measurement model testing

Internal consistency reliability for the five measurement scales was examined using the Cronbach's alpha values. As listed in Table 1, all of these values were greater than the recommended threshold value of 0.7 (Nunnally, 1978).

The item loading, the composite reliability (CR), and the average variance extracted (AVE) were computed (see Table 2) to assess the convergent validity of the questionnaire items (Fornell & Larcker, 1981). The values of the individual factor loadings ranged from 0.539 to 0.953, exceeding the minimum acceptable level of 0.5 proposed by Hair, Anderson, Tatham, and Black (1998) and indicating a well-defined structure. The CR was calculated as indicated by Fornell and Larcker (1981), with results ranging from 0.705 to 0.914, exceeding the critical value of 0.7 (Nunnally, 1978) and indicating adequate CR. The AVE measures the overall amount of variance attributed to the construct relative to the amount of variance attributed to measurement error. The AVE for each construct should be at least 0.4 (Thompson, 2004), at which

Table 1
Descriptive statistics and internal consistency reliability.

Construct and measurement items	Item mean	Std. deviation	Item-total correlation	Internal consistency reliability
Computer self-efficacy (CSE)				
CSE1	3.63	0.741	0.809	0.714
CSE2	3.72	0.723	0.790	
CSE3	3.66	0.764	0.795	
Training satisfaction (TS)				
TS1	4.24	0.888	0.897	0.910
TS2	4.24	0.858	0.944	
TS3	4.16	0.944	0.923	
Test anxiety (TA)				
TA1	2.28	0.665	0.689	0.789
TA2	2.32	0.577	0.665	
TA3	2.28	0.624	0.772	
TA4	2.37	0.589	0.774	
TA5	2.35	0.564	0.792	
CAT attitude (CAT-A)				
CAT-A1	4.03	0.856	0.867	0.806
CAT-A2	4.00	0.837	0.860	
CAT-A3	3.50	0.662	0.831	
CAT performance (CAT-P)				
CAT-P1 (Listening)	70.25	9.976	0.800	0.707
CAT-P2 (Grammar)	73.01	10.770	0.813	
CAT-P3 (Vocabulary)	77.13	9.256	0.771	

Table 2
Convergent validity.

Latent variable	Item	Item loading	CR	AVE
CSE	CSE1	0.718	0.715	0.456
	CSE2	0.671		
	CSE3	0.634		
TS	TS1	0.814	0.914	0.779
	TS2	0.953		
	TS3	0.876		
TA	TA1	0.539	0.797	0.450
	TA2	0.540		
	TA3	0.719		
	TA4	0.736		
	TA5	0.767		
CAT-A	CAT-A1	0.780	0.814	0.593
	CAT-A2	0.743		
	CAT-A3	0.787		
CAT-P	CAT-P1	0.747	0.705	0.446
	CAT-P2	0.617		
	CAT-P3	0.631		

point the variance captured by the construct exceeds the variance due to measurement error. The results were all above 0.4, ranging from 0.446 to 0.779. Overall, the measurement model exhibited appropriate convergent validity.

As shown in Table 3, the square root of AVE shared between a construct and its items (appearing in bold along the diagonal) was greater than the correlations between the construct and any other construct in the model, satisfying Fornell and Larcker's (1981) criteria for discriminant validity. In fact, following the suggestion of a more stringent

Table 3
Discriminant validity.

Construct	CSE	TS	TA	CAT-A	CAT-P
CSE	**0.675**				
TS	0.162	**0.883**			
TA	−0.207	−0.273	**0.671**		
CAT-A	0.181	0.290	−0.165	**0.770**	
CAT-P	0.030	0.070	−0.250	0.185	**0.668**

Note. The diagonal elements in bold (the square root of AVE) should exceed the inter-construct correlations below and across them for adequate discriminant validity.

approach, proposed by Gefen, Straub, and Boudreau (2000), of using the AVEs themselves instead of their square roots across the diagonal renders the same conclusion with respect to discriminant validity. Given the above analysis, the measurements used in this study demonstrated sufficient evidence of construct validity.

3.2. Structural equation model testing

The hypotheses and the paths between the latent construct in M1 were examined with the structural equation model. Fig. 2 showed the completely standardized LISREL path coefficients. The hypothesized SEM model of M1 had a good fit, χ^2/df = 1.601 (χ^2 = 177.684, df = 111), NNFI = 0.962, CFI = 0.969, and RMSEA = 0.0449. The fit indices were within accepted thresholds (Hau, Wen, & Cheng, 2004).

However, the modification indices of M1 explained there was potential misfit in the original research model. A path from training satisfaction to CAT attitude was suggested to add to explore the possibility of any direct effect of training satisfaction on CAT attitude. Fig. 3 showed the first modified research model (M2) and the completely standardized LISREL path coefficients. Compared with M1, the overall fit of M2 was improved significantly, χ^2/df = 1.455 (χ^2 = 160.062, df = 110), NNFI = 0.971, CFI = 0.977, and RMSEA = 0.0391.

Additionally, the path from training satisfaction to CAT performance in M2 was not significant even at the 0.05 level. This result indicated that training satisfaction showed no statistically significant direct effect on CAT performance. Therefore, model modification was conducted to improve the model. To this end, the non-significant path was deleted. The fit indices of the second modified research model (M3) was improved slightly: χ^2/df = 1.443 (χ^2 = 160.162, df = 111), NNFI = 0.972, CFI = 0.977, and RMSEA = 0.0385. The schematic representation of M3 with standardized path coefficients was given in Fig. 4.

Further, SEM analysis of M3 pointed the path from CAT attitude to CAT performance was insignificant. So the insignificant path was also deleted. According to the results of SEM analysis of present model, there was no more modification information. In this case, the last research model (M4) was retained. Compared with the other structural equation models, M4 not only had the best fit to present data: χ^2/df = 1.432 (χ^2 = 160.398, df = 112), NNFI = 0.973, CFI = 0.977, and RMSEA = 0.038, but also was the most parsimonious model. Fig. 5 illustrated the diagram of M4 including the standardized estimates.

Fig. 2. SEM analysis of M1.

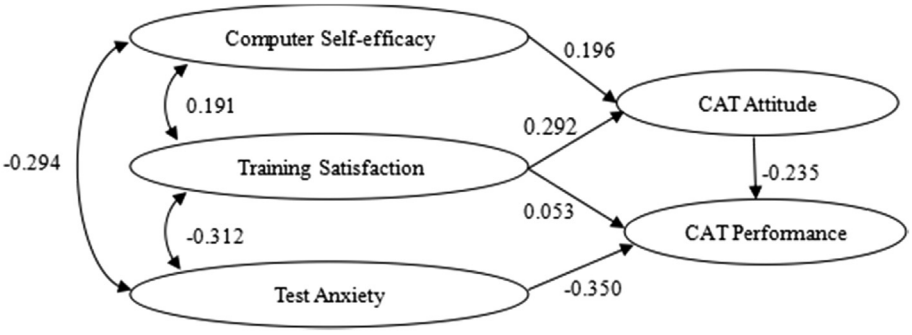

Fig. 3. SEM analysis of M2.

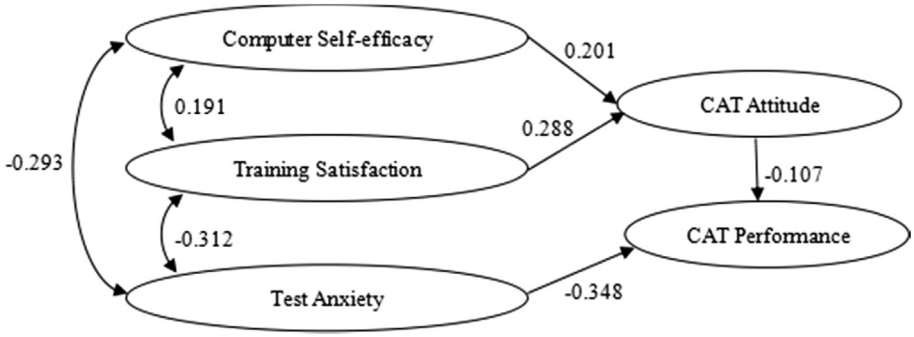

Fig. 4. SEM analysis of M3.

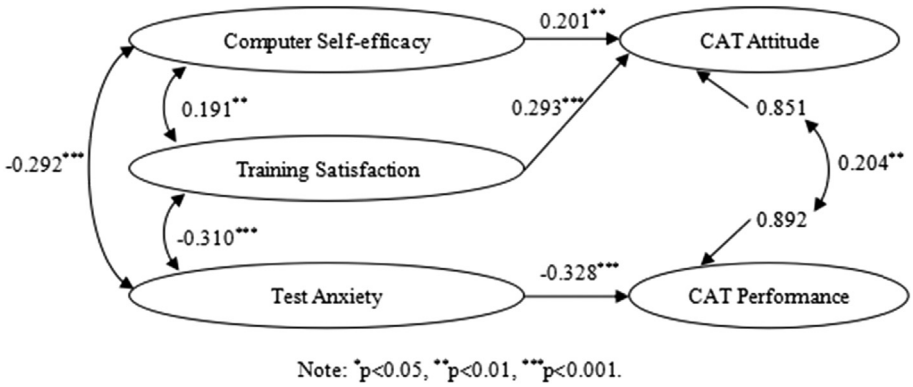

Fig. 5. SEM analysis of M4.

4. Discussion

In the last research model, significant positive paths were obtained leading from computer self-efficacy and training satisfaction to their destination of CAT attitude. Test anxiety had a significant negative effect on CAT performance. The paths were all found to be significant at least at the level of 0.01.

The first hypothesis, which stated that higher CAT attitude would exhibit higher CAT performance, was not supported. The result might be explained in the light of following arguments. CAT performance in the present study was mainly influenced by test-takers' English ability. However, CAT attitude was the feelings of test-takers towards CAT. Higher CAT attitude only meant that test-takers tended to take the test mode. But it was unrelated to the higher or lower level of English ability. Thus, it was improper for researchers to link test-takers' performance to their attitude towards test mode either in CAT or P&P. Furthermore, there was significant correlation between the residual variances of CAT attitude and CAT performance. For a deeper understanding of the correlation, it would require a further study to investigate the homogeneity between CAT attitude and CAT performance.

As expected, the results of this study provided strong support for the second hypothesis and confirmed that the positive effect of computer self-efficacy on CAT attitude was significant. It was in line with previous research which argued that individuals with high perceived computer self-efficacy were more inclined to using computer supported education (Celik & Yesilyurt, 2013; Pellas, 2014). Indeed, with the fact that computers and Internet access were more and more cheap and had been available in recent years, a degree of familiarity with standard computer software packages was becoming a basic requirement for students. The developing trend would not only be in favor of increasing test-takers' computer self-efficacy, but also be helpful improving test-takers' CAT attitude.

The third hypothesis examined the links between training satisfaction and CAT performance. The present study didn't find the significant influence of training satisfaction on CAT performance. The result seemed to be inconsistent with the findings obtained in Ortner and Caspers (2011) suggesting that informing test-takers about the mechanisms of adaptive testing led to higher scores. One possible explanation for the different result might be the different experimental procedure. In the study of Ortner and Caspers (2011), the sample was divided into two parts. Half of the sample who worked on the adaptive version received information about how adaptive tests work, and the other half of the sample received standard instruction without

information on adaptive testing. There were significant different scores between the two parts. However, in the present study, all of the test-takers took part in CAT training. The SEM analysis of M4 revealed that, when test-takers knew adaptive testing mechanism by training, their training satisfaction would influence CAT attitude rather than CAT performance.

With regard to the fourth hypothesis, it appeared that test anxiety negatively and significantly affected CAT performance. The result of the present study was not only in accordance with most prior CAT researches (Kim & McLean, 1994; Ortner & Caspers, 2011), but also in agreement with those P&P researches (Chapell et al., 2005; Iroegbu, 2013; Rezazadeh & Tavakoli, 2009; Trifoni & Shahini, 2011). In other words, no matter which test mode was carried out, the more test anxiety test-takers felt, the worse they would perform. In addition, Ortner and Caspers (2011) also indicated that, compared CAT with P&P, CAT item difficulty increased more quickly and success probability decreased more quickly, and thus increased state anxiety, especially for individuals possessing high levels of trait test anxiety. In view of this finding, test anxiety was assumed to have negative effect on CAT attitude. However, the SEM analysis of M4 in the present study found that there was no causal relationship between test anxiety and CAT attitude. The difference between the two research procedures was that there was CAT training in this study. In order to examine whether or not CAT training led to the different finding, a further study of the relationship among test anxiety, CAT training and CAT attitude would be conducted.

H5, H6, and H7 postulated that computer self-efficacy, training satisfaction, and test anxiety were inter-correlated. All of these hypotheses were supported by this study. The results showed a significant negative correlation between computer self-efficacy and test anxiety. The significant predicting functions from computer self-efficacy and test anxiety on training satisfaction were also revealed, specifically, the higher test-takers' computer self-efficacy, the higher training satisfaction they would feel, and the higher test-takers' test anxiety, the lower training satisfaction they would feel.

5. Conclusion and implication

The present study tested a CAT model by exploring the causal paths among a series of individual variables. The model showed good fit to the dataset regarding the evaluated variables. Besides, compared with previous researches, it was the first time to explore the effect of computer self-efficacy in the CAT model.

The main purpose of the present study was to assess the test fairness of CAT through examining the effect of different individual variables on CAT performance. Test fairness was required to remove the variance, which was attributable to individual variables that were irrelevant to measurement of the construct of interest, from test scores as far as possible (Helms, 2004). Among the variables tested in the present study, CAT attitude, computer self-efficacy and training satisfaction proved to have insignificant influence on CAT performance, and only test anxiety had significant negative influence on CAT performance. So it could be seen CAT might produce an unfair disadvantage for test-takers with higher test anxiety.

In addition, because of the existing clear difference between CAT and P&P, though CAT had many advantages, people showed different CAT attitudes. Some people held the opinion that CAT, as a test mode, wasn't the primary choice. This study pointed that increasing test-takers' computer self-efficacy or training satisfaction could improve their CAT attitude evidently, especially training satisfaction could show this effect in a short time. Thus, how to design training content as well as training method to improve test-takers' training satisfaction should be the focus for CAT researchers.

What's more, there were significant differences in information technology facilities in different areas, such as rural and urban area. As a result, during the generalization of CAT, people often concerned that test-takers would vary in computer self-efficacy, training satisfaction and CAT attitude (Saleem, Beaudry, & Croteau, 2011; Scott & Walczak, 2009), which might cause the difference in test-takers' CAT performance. Therefore, CAT seemed to be unfair for the test-takers from poorly-equipped areas. However, the final model in this paper indicated that such concern could be negligible.

6. Limitation and future research

One limitation of the present study was that it was the first time for test-takers to take CAT. Furthermore, all of the examinees were unfamiliar with CAT. Kravitz, Stinson, and Chavez (1996) had ever examined participant reactions to a variety of different selection and promotion procedures such as interviews, work samples, drug tests, etc. The results were found that the more experience one had with a selection procedure, the more positively the procedure was evaluated. Thus, if test-takers had more opportunity to take CAT, their CAT attitude would be improved. In future research, the moderating role of CAT attitude in the model could be addressed more sufficiently with a longitudinal design.

Another limitation lied in sample bias. Participants in this study were from urban high school in China, who were more familiar with information technology and easier to adopt newborn things compared with students from rural high school (Cai, 2014; Li, 2013; Wang, 2013b). As a result, in order to make the conclusions of similar study be more persuasive, sample from rural areas must be taken into consideration in the future.

The third limitation was that only three individual characteristics were included in the simple causal model of this study as exogenous latent variables. In fact, other individual characteristics, such as the average response time/number of items answered by the students, might also have relationship with CAT attitude or CAT performance. For instance, higher ability examinees usually spent significantly more average time on CAT questions than did their counterparts with lower ability (Chang, Plake, & Ferdous, 2005). So, a complicated causal model which included more latent variables should be built in future research. Undoubtedly, more suggestions which were beneficial to the implementation of CAT would be revealed by analyzing the complicated model.

Appendix

Table 4
Measurement items.

Construct	Measure
Computer self-efficacy (CSE)	
CSE1	I could complete the job using the software package if I had used similar packages before this one to do the same job.
CSE2	I could complete the job using the software package if I had a lot of time to complete the job for which the software was provided.
CSE3	I could complete the job using the software package if I could get help from the Internet.
Training satisfaction (TS)	
TS1	In my opinion the planned objectives were met.
TS2	The training was realistic and practical.
TS3	The training received is useful for my specific job.
Test anxiety (TA)	
TA1	During tests I feel very tense.
TA2	I wish examinations did not bother me so much.
TA3	I seem to defeat myself while working on important tests.
TA4	I feel very panicky when I take an important test.
TA5	During examinations I get so nervous that I forget facts I really know.
CAT attitude (CAT-A)	
CAT-A1	I think adaptive testing can obtain accurate information about test-takers' abilities.
CAT-A2	I think adaptive testing is fair for test-takers.
CAT-A3	I like this type of test.

References

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*(1), 117–148.

Bolliger, D. U., & Halupa, C. (2012). Student perceptions of satisfaction and anxiety in an online doctoral program. *Distance Education, 33*(1), 81–98.

Cai, Z. L. (2014). An investigation and study on the differences between urban and rural students' scientific literacy. *Education Teaching Forum, 7*(43), 91–93.

Celik, V., & Yesilyurt, E. (2013). Attitudes to technology, perceived computer self-efficacy and computer anxiety as predictors of computer supported education. *Computers & Education, 60*(1), 148–158.

Chang, S. R., Plake, S. B., & Ferdous, A. A. (2005). Response times for correct and incorrect item responses on computerized adaptive tests. In *Paper presented at the 2005 annual meeting of the American Educational Research Association (AERA), Montréal, Canada*.

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., et al. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97*(2), 268–274.

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.

Compeau, D., & Higgins, C. (1995). Computer self-efficacy: development of a measure and initial test. *MIS Quarterly, 19*(2), 189–202.

Davis, R. M. (2012). Adaptive testing evolves to assess common-core skills. *Education Week*, (1), 12–16.

Farooqi, Y. N., Ghani, R., & Spielberger, C. D. (2012). Gender differences in test anxiety and academic achievement of medical students. *International Journal of Psychology and Behavioral Sciences, 2*(2), 38–43.

Finkelman, M. D., Kim, W., Weissman, A., & Cook, R. J. (2014). Cognitive diagnostic models and computerized adaptive testing: two new item-selection methods that incorporate response times. *Journal of Computerized Adaptive Testing, 2*(4), 59–76.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50.

Fritts, B. E., & Marszalek, J. M. (2010). Computerized adaptive testing, anxiety levels, and gender differences. *Social Psychology of Education, 13*(3), 441–458.

Gefen, D., Straub, D. W., & Boudreau, M. C. (2000). Structural equation modeling and regression: guidelines for research practice. *Communications of the AIS, 4*(7), 2–77.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis with readings*. Upper Saddle River, NJ: Prentice-Hall.

Hau, K., Wen, Z., & Cheng, Z. (2004). *Structural equation model and its applications*. Beijing: Educational Science Publishing House.

He, W., Diao, Q., & Hauser, C. (2014). A comparison of four item-selection methods for severely constrained CATs. *Educational and Psychological Measurement, 74*(4), 677–696.

Helms, J. E. (2004). Fair and valid use of educational testing in grades K-12. In J. E. Wall, & G. R. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 81–88). Greensboro, NC: CAPS Press.

He, W., & Reckase, M. D. (2014). Item pool design for an operational variable-length computerized adaptive test. *Educational and Psychological Measurement, 74*(3), 473–494.

Iroegbu, M. N. (2013). Effect of test anxiety, gender and perceived self-concept on academic performance of Nigerian students. *International Journal of Psychology and Counselling, 5*(7), 143–146.

Johnson, R. D., Gueutal, H., & Falbe, C. M. (2009). Technology, trainees, metacognitive activity and e-learning effectiveness. *Journal of Managerial Psychology, 24*(6), 545–566.

Johnson, D. F., & Mihal, W. L. (1973). Performance of blacks and whites in computerized versus manual testing environments. *American Psychologist, 28*(8), 694–699.

Jung, H. (2014). Ubiquitous learning: determinants impacting learners' satisfaction and performance with smartphones. *Language Learning and Technology, 18*(3), 97–119.

Kim, J. W., Han, D. H., Lee, Y. S., Min, K. J., Park, J. Y., & Lee, K. (2013). The effect of depression, anxiety, self-esteem, temperament, and character on life satisfaction in college students. *Journal of Korean Neuropsychiatric Association, 52*(3), 150–156.

Kim, J. G., & McLean, J. E. (1994). The relationships between individual difference variables and test performance in computerized adaptive testing. In *Paper presented at the annual meeting of the Mid-South Educational Research Association, Nashville, TN, November 9-11, 1994*.

Kravitz, D. A., Stinson, V., & Chavez, T. L. (1996). Evaluations of tests used for making selection and promotion decisions. *International Journal of Selection and Assessment, 4*(1), 24–34.

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.

Li, B. P. (2013). Analysis on the differences between the primary and middle school students' information learning ability. *Journal of the Chinese Society of Education, 36*(3), 20–23.

McGuire, W. J. (1985). Attitudes and attitude change. In G. Lindzey, & E. Aronson (Eds.), *Handbook of social psychology* (vol. 2). New York: Random House.
Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
Ortner, M. T., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment, 27*(3), 157–163.
Ortner, M. T., Weißkopf, E., & Koch, T. (2014). Higher ability students' motivational experiences during adaptive achievement testing. *European Journal of Psychological Assessment, 30*(1), 48–56.
Papanastasiou, E., & Zembylas, M. (2002). The effect of attitudes on science achievement: a study conducted among high school pupils in Cyprus. *International Review of Education, 48*(6), 469–484.
Paul, R., Hauser, R. D., & Bradley, J. H. (2007). The relationship between individual differences, culture, anxiety, computer self-efficacy and user performance. *International Journal of Information Systems and Change Management, 2*(2), 125–138.
Pellas, N. (2014). The influence of computer self-efficacy, metacognitive self-regulation and self-esteem on student engagement in online learning programs: evidence from the virtual world of second life. *Computers in Human Behavior, 35*, 157–170.
Rezazadeh, M., & Tavakoli, M. (2009). Investigating the relationship among test anxiety, gender, academic achievement and years of study: a case of Iranian EFL university students. *English Language Teaching, 2*(4), 68–74.
Saleem, H., Beaudry, A., & Croteau, A. M. (2011). Antecedents of computer self-efficacy: a study of the role of personality traits and gender. *Computers in Human Behavior, 27*, 1922–1936.
Scott, E. J., & Walczak, S. (2009). Cognitive engagement with a multimedia ERP training tool: assessing computer self-efficacy and technology acceptance. *Information & Management, 46*, 221–232.
Shen, C. W., Wu, Y. C., & Lee, T. C. (2014). Developing a NFC-equipped smart classroom: effects on attitudes toward computer science. *Computers in Human Behavior, 30*, 731–738.
Singh, A., Bhadauria, V., Jain, A., & Gurung, A. (2013). Role of gender, self-efficacy, anxiety and testing formats in learning spreadsheets. *Computers in Human Behavior, 29*(3), 739–746.
Spielberger, C. D., Gonzalez, H. P., Taylor, C. J., Anton, E. D., Algaze, B., Ross, G. R., et al. (1980). *Manual for the test anxiety inventory ("Test attitude inventory")*. Redwood City, CA: Consulting Psychologists Press.
Spielberger, C. D., & Vagg, P. R. (1995). *Test anxiety, a transactional process model*. Washington, DC: Taylor and Francis.
Taylor, J., & Deane, F. P. (2002). Development of a short form of the test anxiety inventory (TAI). *The Journal of General Psychology, 129*(2), 127–136.
Tello, F., Moscoso, C. S., Garcia, B. I., & Chaves, S. S. (2006). Training satisfaction rating scale. *European Journal of Psychological Assessment, 22*(4), 268–279.
Thompson, B. (2004). *Exploratory and confirmatory factor Analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.
Tonidandel, S., & Quinones, M. A. (2000). Psychological reactions to adaptive testing. *International Journal of Selection and Assessment, 8*(1), 7–15.
Trifoni, A., & Shahini, M. (2011). How does exam anxiety affect the achievement of university students? *Mediterranean Journal of Social Science, 2*(2), 93–100.
Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
Wang, C. (2013a). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement, 73*(6), 1017–1035.
Wang, Z. L. (2013b). The strategy and path of educational information in China. *Modern Distance Education, 32*(4), 62–69.
Wang, C., Chang, H. H., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement, 37*(2), 99–122.
Weiss, D. J. (1975). Adaptive testing research at Minnesota: overview, recent results, and future directions. In , *Professional series 75–6Proceedings of the first conference on computerized adaptive testing*. Washington, DC: Personnel Research and Development Center, U.S. Civil Service Commission.
Weiss, D. J., & Betz, N. E. (1973). *Ability Measurement: Conventional or adaptive?* (Research Report 73–1). Minneapolis, MN: University of Minnesota, Department of Psychology.
Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: theory and applications. *Psychometrika, 77*(3), 495–523.
Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement, 37*(1), 3–23.
Yukselturk, E., & Bulut, S. (2007). Predictors for student success in an online course. *Educational Technology & Society, 10*(2), 71–83.
Zeidner, M. (1998). *Test Anxiety: The State of the Art*. New York: Springer.