Mobile Testing for Authentic Assessment in the Field:Evaluation from Actual Performances

Yoshimitsu Miyasawa
University of Electro-Communications, Tokyo, Japan
Email: miyasawa@ai.is.uec.ac.jp

Maomi Ueno
University of Electro-Communications, Tokyo, Japan
Email: ueno@ai.is.uec.ac.jp

***Abstract*—We have developed a mobile testing system using adaptive testing for assessing learning at museums, parks, and other sites in the field. Adaptive testing is a form of computer-based testing that progressively estimates an examinee's ability from his/her answer history and uses that ability to present test items making ability estimation even more accurate. Field-testing, however, requires activities such as observing and searching at specific positions within a site, which requires the learner to move about to get to those positions. Moreover, the time that can be spent taking such an on-site test is usually limited, which means that the test may end before a sufficient number of test items can be answered thereby decreasing the accuracy of ability estimation. In response to these issues, we formalize for field-testing purposes an optimization problem called the traveling purchaser problem (TPP) that incorporates graph theory and propose an adaptive testing system using TPP.**

I. Introduction

Knowledge does not exist on its own; rather, it is embedded in situations [1]. It is known, moreover, that knowledge is acquired in conjunction with past experiences [2]. These observations suggest that testing should be embedded in situations to authentically assess learning. Recent advances in mobile technologies are making it possible to estimate ability in a manner not possible by paper-based testing. Specifically, they are making it possible to perform assessments that require actions like observing and searching in the field such as at museums as opposed to tests that simply assess knowledge related to facts and procedures [3].

Taking, for example, a testing system using the Global Positioning System (GPS) has been developed for administering tests that require actions like observing and searching in the field[3]. This system takes into account the fact that test items require certain actions at a specific position and therefore identify the examinee's present position to present test items corresponding to that position. E-learning and learning using mobile devices[4], [5], [6], [7], systems have also been developed to support these forms of learning in the field in terms of formative evaluation and self/peer assessment [8], [9], [10]. However, the test items presented by this system are fixed, that is, the same test items are presented to all examinees.

A more effective presentation format has recently been achieved through Computerized Adaptive Testing (CAT). CAT progressively estimates the examinee's ability from his/her answer history and uses an item bank to present test items that maximize the amount of information with regard to that ability[11], [12], [13], [14], [15]. Furthermore, by selecting test items most applicable to ability under a time-limit constraint, the accuracy of estimating ability can be improved for tests having a time limit [16], [17]. Improvement in the accuracy of estimating the examinee's ability is one advantage that can be expected from CAT, and incorporating CAT in mobile testing systems should be able to improve the accuracy of ability estimation in anywhere/anytime testing [18], [19].

Testing in the field, however, requires actions like observing and searching at specific positions within a certain site, which means that the examinee must move about to get to those positions. In other words, positions at which the examinee must respond to test items are scattered throughout a site, which means that wasted time from unnecessary back-and-forth movements can be incurred. Moreover, as the time that can be spent for taking a test is generally limited, there is always the possibility that the test will end before a sufficient number of test items have been answered thereby decreasing the accuracy of ability estimation.

The purpose of this study is to make tests in the field more efficient and improve the accuracy of estimating an examinee's ability. Specifically, we propose a CAT system using the traveling purchaser problem (TPP), which is an optimization problem combined with graph theory.

The TPP is defined as follows[20]. Let nodes and edges within a graph denote stores and distance traveled, respectively. Each store sells products that need to be purchased but the number of products and their prices differ from store to store. The task here is to find a route that returns to the purchaser's point of departure minimizing the total cost of products and distance travelled. For the purposes of our study, we change products and shops defined in TPP to test items and the positions where those test items are presented, respectively, with the aim of finding the optimal route in a mobile test. In TPP, however, the number of products is given as a constraint, but since it is our desire to give time as a constraint in our study, we cannot use TPP in its existing form.

We therefore propose TPP having a time-limit constraint and propose a CAT system using this modified form of TPP as an optimization problem. The advantage of this approach is that we can raise the efficiency of testing that considers movement in the field and therefore improve the accuracy of estimating the examinee's ability.

In this paper, we also report on simulations and experiments that show the proposed system to be more accurate in measuring performance compared to previous systems. In addition, we show more detailed analysis compared to our own previous work[21].

II. Test Theory

Most forms of CAT make use of a type of test theory called Item Response Theory (IRT) [11], [12], [13], [14], [15]. This is a mathematical model that enables examinees to be compared on the same scale even in they take different tests.

We denote the response of examinee j to item i as u_{ij}, which we define as follows:

$$u_{ij} = \begin{cases} 1 & \text{If } j\text{-th examinee answers} \\ & i\text{-th item correctly} \\ 0 & \text{Other Cases} \end{cases}.$$

In IRT, the ability of examinee j is expressed as ability $\theta_j \in (-\infty, \infty)$ and the probability of examinee j responding correctly to item i is expressed as a function of θ_j. For example, the widely used two-parameter logistic model expresses this probability as follows:

$$P(u_{ij} = 1|\theta_j) = \frac{1}{1 + \exp\{-1.7a_i(\theta_j - b_i)\}} \quad (1)$$

In the equation, a_i denotes the discrimination parameter that indicates the degree to which item i discriminates with respect to ability θ_j, and b_i denotes the difficulty parameter that, as the name implies, indicates item difficulty.

For each item, the scale representing the accuracy of estimating the ability is called "item information," which is generally expressed as Fisher information defined as follows:

$$I(\theta_j|a_i, b_i) = 1.7^2 a_i^2 P(\theta_j|a_i, b_i)(1 - P(\theta_j|a_i, b_i)) \quad (2)$$

The reciprocal of Fisher information converges asymptotically to the dispersion (square of the error) of the estimated value of the ability. Thus, in CAT, test items with higher item information are successively presented.

III. CAT Incorporating TPP with a Time-limit Constraint

With the aim of using TPP with time as a constraint for testing conducted in the field, we formalize this type of TPP as an optimization problem in the following way. Let the set of positions that present test items be denoted as $S := \{v_1, \ldots, v_n\}$ and the set of all items as $K := \{p_1, \ldots, p_m\}$. Let the set of all positions be denoted as $S_0 := \{S \cup o\}$, where o is the point of departure. Graph $G = (V, E)$ is an undirected graph, where $V := S_0$ represents the set of nodes and $E := \{[v_i, v_j] : v_i, v_j \in S, i < j\}$ the set of edges. Let the item information of test item p_k be denoted as b_k, the time required to answer test item p_k (required response time) as t_k, and the travel time between positions v_i, v_j as d_{ij}. T is the test time limit. The order of presenting items is called a route. For a certain route, if item p_k of position v_i is included in the route, $z_{ik} = 1$, and if not, $z_{ik} = 0$. Furthermore, if the route between positions v_i and v_j is included, $x_{ij} = 1$, and if not, $x_{ij} = 0$. The optimal route can now be found from the following optimization problem:

$$\text{Maximize } w = \sum_{v_i \in S} \sum_{p_k \in K} b_k z_{ik} \quad (3)$$

subject to

$$\sum_{v_i \in S} \sum_{p_k \in K} t_k z_{ik} + \sum_{(i,j) \in L} d_{ij} x_{ij} < T \quad (4)$$

(3) is the objective function maximizing item information and (4) is the constraint that keeps the total sum of travel time between items and required response times of each test item within the designated test time limit. Solving this optimization problem finds a route with a maximum amount of item information within the time limit. However, there is no guarantee that this optimal route will minimize travel time. In other words, there is still the possibility that a roundabout way will be followed even if the route is within the time limit. We therefore add to the objective function a travel-time term having the constant $D << 1$ as follows:

$$\text{Maximize } w = \sum_{v_i \in S} \sum_{p_k \in K} b_k z_{ik} - D \sum_{(i,j) \in L} d_{ij} x_{ij} \quad (5)$$

subject to

$$\sum_{v_i \in S} \sum_{p_k \in K} t_k z_{ik} + \sum_{(i,j) \in L} d_{ij} x_{ij} < T \quad (6)$$

Solving this optimization problem determines which test items to present to the examinee.

The algorithm for CAT incorporating this optimization problem is shown in Figure 1. Since item information represented by (5) differs according to the ability of the examinee, the optimization problem of (5) is iteratively estimated after updating the examinee's ability. This algorithm begins by setting the initial value of the examinee's ability to 0. It then estimates item information for all test items based on the examinee's ability and solves the optimization problem thereby finding the optimal route and selecting the next test item to be presented. The examinee now selects an answer from the options (multiple choice) presented for that test item. The system then automatically determines whether that answer is correct or incorrect and re-estimates the examinee's ability using that result. The above process is repeated to estimate the ability of the examinee.

The computational complexity of this optimization problem is quite high, which means that iteratively solving the problem based on the examinee's answer to select an optimal test item from the set of all test items is difficult. For this reason, we divide the test item presentation area into multiple sections to lower the computational complexity. In this study, the number of items in each section is set to no more than 20. Simulations have shown that this number enables a computer to find an optimal solution within three seconds.

IV. Simulation Experiment

In this section, we describe a simulation experiment that we conducted to confirm the validity of the proposed method. This experiment compared total item information with respect to true ability for various methods, namely, the proposed method, time-constrained CAT[16], [17], general CAT [11], [12], [13], [14], [15], and test item selection using a random number. Characteristics of the test items used in this simulation were generated from random numbers based on IRT.

Specifically, the true values of a_i, b_i, t_i, x_i, and y_i of item i were generated from the following random numbers:

- No. of items: 20
- $a_i \sim N(1.7, 0.5^2)$

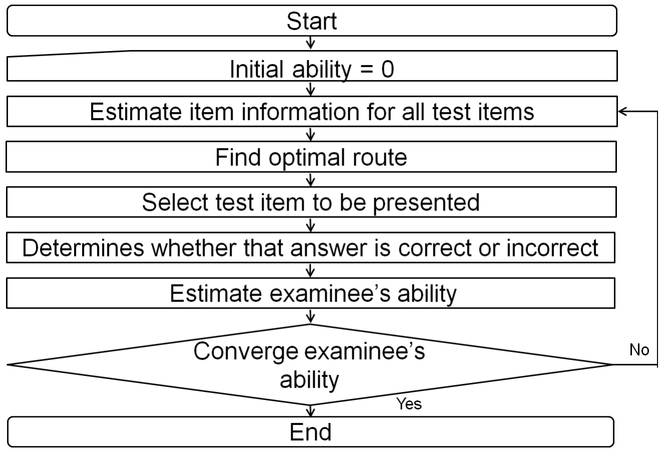

Fig. 1. Adaptive testing algorithm for mobile testing

- $b_i \sim N(0, 0.5^2)$
- $t_i \sim N(30, 10^2)$
- position$x \sim U(0, 300)$
- position$y \sim U(0, 300)$

To simplify calculations here, we set the value of the distance between position (x_i, y_i) of item i to position (x_i+1, y_i+1) of item $i+1$ to the travel time [s] from item i to item $i+1$.

The experiment compared the average total item information for 100 examinees for each of the true abilities of -2.0, -1.0, 0.0, 1.0 and 2.0 and time-limit constraint values of 600 [s] and 1000 [s].

Simulation results are listed in Table IIIIII. These results showed that item information achieved by the proposed method was high relative to the other methods for each true ability. This says, in other words, that the accuracy of estimating ability was high when using the proposed method compared to the other methods. These results also showed that the travel time spent by the proposed method was short relative to the other methods for each true ability, although Number of test items provided by the proposed method was large. Namely, the order of moving spent to a test was the optimal when using the proposed method compared to the other methods. Yet, for a time limit of 600 [s], there were times when item information by the proposed method was lower than that by time-constrained CAT.

We next clarify the cause of this phenomenon. Figure 2 shows item information (vertical axis) versus time limit (horizontal axis) for the proposed method and time-constrained CAT. These results show that item information by the proposed method was low for short time limits. We therefore analyzed the test items presented to examinees for a short time limit. The test items presented by the proposed method are shown in Figure 4 and those by time-constrained CAT are shown in Figure 5. Examining these results, it can be seen that the test items initially presented by the proposed method not only had low item information but also had long traveling time. This came about because the value of the examinee's ability in the early stage had not yet converged, i.e., test item selection was

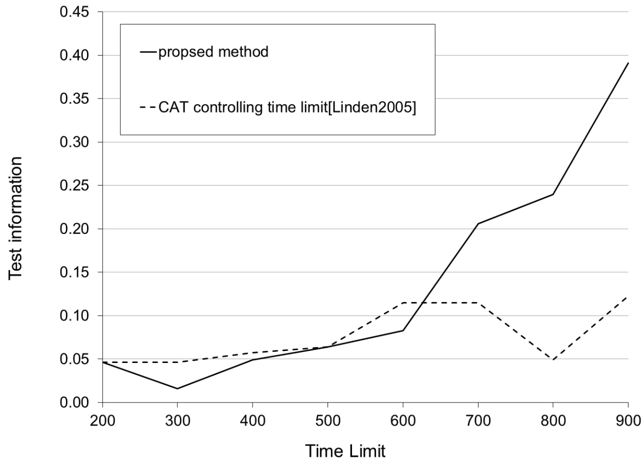

Fig. 2. Adaptive testing algorithm for mobile testing

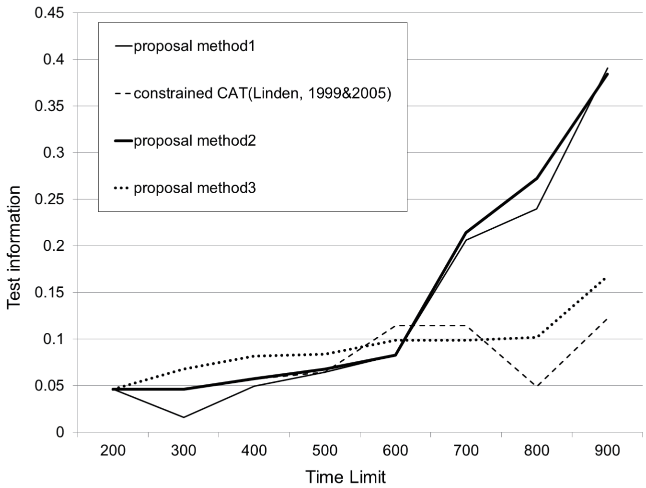

Fig. 3. Analysis of proposed method

being performed according to an ability divergent from the examinee's true ability resulting in the selection of test items with low item information.

To rectify this problem, we incorporated in our proposed method a rule stating that the initially presented test item must be the one closest to the current position. In Figure 3, proposed method 2 applies this rule of presenting the test item closest to the current position only to the first initial item, while proposed method 3 applies this method to the first seven initial test items. These results show that item information by proposed method 2 was not low compared with that of time-constrained CAT, but that item information by proposed method 3 was low for long time limits. What this means is that adjusting the initial test item does not decrease item information for short time limits, or in other words, that adjusting the initial test item can improve the accuracy of estimating ability regardless of the length of the time limit.

V. Mobile Testing System

We here describe the mobile testing system that we developed as part of this study. System configuration is shown in

TABLE I. AVERAGE VALUES OF ITEM INFORMATION

Time Limit[s]	True Ability	Proposed Method	CAT constrained -Time Limit	CAT	Random
600	-2.0	0.0828 (0)	0.114 (0.00263^2)	0.00772 (0)	0.0599 (0.0443^2)
	-1.0	0.956 (0.0578^2)	1.11 (0.253^2)	0.180 (0)	0.635 (0.275^2)
	0.0	2.86 (0.151^2)	1.43 (0.214^2)	0.888 (0)	1.38 (0.574^2)
	1.0	1.17 (0.153^2)	0.786 (0.122^2)	0.0945 (0)	0.457 (0.247^2)
	2.0	0.240 (0.0366^2)	0.0434 (0)	0.00384 (0)	0.0688 (0.0698^2)
1000	-2.0	0.415 (0)	0.123 (0.00572^2)	0.0791 (0.0146^2)	0.109 (0.0672^2)
	-1.0	3.00 (0.278^2)	1.44 (0.334^2)	1.09 (0.408^2)	1.04 (0.440^2)
	0	5.50 (0.348^2)	2.64 (0.422^2)	2.38 (0.440^2)	2.33 (0.578^2)
	1.0	2.45 (0.117^2)	1.22 (0.199^2)	1.15 (0.126^2)	0.917 (0.318^2)
	2.0	0.299 (0.0142^2)	0.0781 (0)	0.0778 (0.00288^2)	0.120 (0.0788^2)

TABLE II. AVERAGE VALUES OF NUMBER OF TEST ITEMS

Time Limit[s]	True Ability	Proposed Method	CAT constrained -Time Limit	CAT	Random
600	-2.0	5.0 (0)	3.0 (0)	1.0 (0)	2.7 (0.985^2)
	-1.0	5.0 (0)	2.95 (0.218^2)	1.0 (0)	3.0 (0.949^2)
	0.0	5.0 (0)	2.57 (0.495^2)	1.0 (0)	2.80 (1.09^2)
	1.0	5.0 (0)	2.07 (0.255^2)	1.0 (0)	2.76 (0.981^2)
	2.0	5.0 (0)	2.0 (0)	1.0 (0)	2.73 (0.999^2)
1000	-2.0	9.0 (0)	4.03 (0.171^2)	2.8 (0.392^2)	4.88 (1.09^2)
	-1.0	9.3 (0.720^2)	4.22 (0.414^2)	2.65 (0.931^2)	4.89 (1.05^2)
	0	9.5 (0.608^2)	4.05 (0.497^2)	3.71 (0.535^2)	4.79 (1.10^2)
	1.0	8.49 (0.831^2)	3.30 (0.686^2)	3.02 (0.140^2)	4.93 (1.09^2)
	2.0	8.01 (0.0995^2)	3.04 (0.280^2)	3.0 (0)	5.01 (1.00^2)

TABLE III. AVERAGE VALUES OF TRAVEL TIME

Time Limit[s]	True Ability	Proposed Method	CAT constrained -Time Limit	CAT	Random
600	-2.0	472 (0)	507 (0.0995^2)	716 (0)	492 (54.2^2)
	-1.0	470 (7.97^2)	508 (7.48^2)	716 (0)	486 (53.3^2)
	0.0	460 (16.0^2)	521 (17.3^2)	716 (0)	487 (59.7^2)
	1.0	440 (11.0^2)	539 (8.93^2)	716 (0)	495 (56.8^2)
	2.0	435 (5.95^2)	541 (0)	716 (0)	493 (61.5^2)
1000	-2.0	752 (0)	754 (16.9^2)	737 (17.5^2)	820 (44.1^2)
	-1.0	783 (23.3^2)	780 (39.9^2)	762 (42.7^2)	815 (56.3^2)
	0	701 (11.7^2)	830 (20.4^2)	822 (24.8^2)	827 (46.6^2)
	1.0	726 (11.2^2)	863 (12.5^2)	851 (30.3^2)	819 (48.1^2)
	2.0	729 (1.19^2)	868 (4.48)	868 (6.96^2)	821 (45.8^2)

Figure 6. This system employs an Android mobile terminal and consists of a navigation function and an item view function. There is also an item selection function on the system's Web server to display a selected test item on the mobile device.

The navigation function displays on a map the examinee's present location and the positions presenting test items. The system obtains the examinee's present location using a GPS function and the examinee's orientation by a geomagnetic

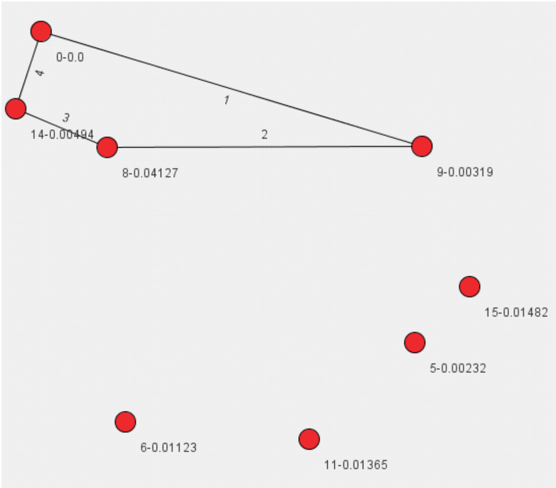

Fig. 4. Route of proposed method

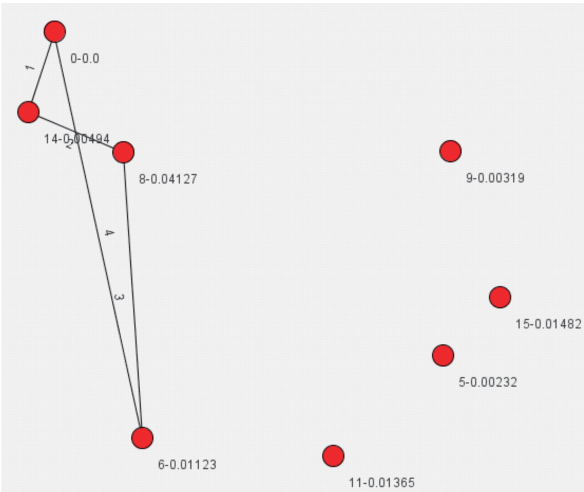

Fig. 5. Route of time-constrained CAT

sensor. The map is achieved using Google Maps provided from the Android Maps API. An example of using this function is shown in Figure 7. In the map, a blue circle indicates examinee's present location and purple marker indicates a position presenting test items. The button at the top of the screen switches the display to the item-viewing screen.

The item view function displays test items to the examinee as shown in Figure 8. All presented test items are in multiple-choice format. The item selection function selects an test item from an item bank based on the proposed CAT system incorporating TPP with a time-limit constraint.

VI. Evaluation of the System

In this section, we describe a experiment that we conducted at a temple site within the Tokyo to evaluate the validity of the proposed system. This site contains a number of national historic landmarks and Buddha statues within an area of about six hectares.

In this experiment, we conducted a test using the proposed method and a test using time-constrained CAT ([16], [17]). For

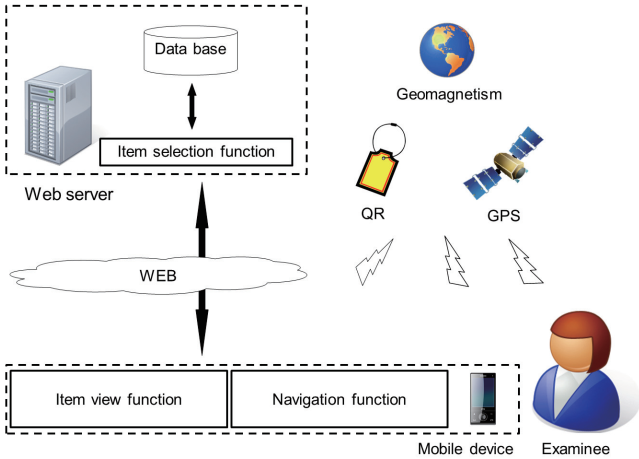

Fig. 6. System configuration

Fig. 7. Screen shot of navigation function

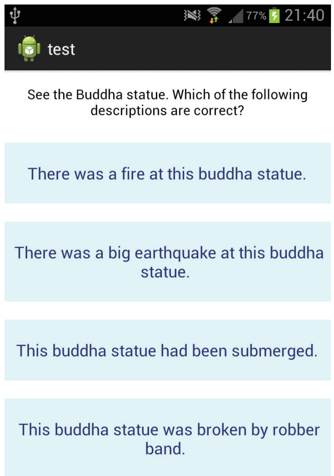

Fig. 8. Screen shot of adaptive testing

each, we analyzed the accuracy of estimating examinee ability and the results of questionnaires. Different two groups of five students in the same university examined each test. They began using the system after being briefed on the experiment and given instructions on using the system. Presented test items asked the subject's knowledge in relation to the target temple site. The test items were written so that actions like observing and searching would be required at specific positions within the site. Item characteristics a_i, b_i, t_i, x_i, and y_i were estimated beforehand. The number of test items was 80.

The questionnaire given to subjects after each test consisted of the two following questions. Subjects were asked to reply to these questions on a six-level basis (6: very true, 5: true, 4: more likely true, 3: more likely false, 2: false, 1: completely false).

1) Travel time was not overly long but appropriate

TABLE IV. THE RESULT OF THE EXPERIMENT (** SIGNIFICANCE LEVEL OF 1%; * SIGNIFICANCE LEVEL OF 5%)

	Proposed Method	Time-constrained CAT
Item information**	4.24(0.236)	2.03(0.253)
Travel time*	110(787)	238(8833)
Number of moves*	4.0(0.5)	5.8(1.7)
Number of test items**	18.2(13.7)	9.0(4)
Question 1**	4.8(0.2)	3.4(0.3)
Question 2*	4.6(0.3)	3.2(0.7)

compared to test item response time.

2) The order of moving from one position to another was the optimal.

In addition, the system estimated travel time and the number of moves.

Results of the experiment are listed in Table IV. Item information was computed for each subject using the ability. The values shown for item information in the table indicates the average value and the variance (in parentheses) of item information over all subjects. The table also lists the average value and the variance (in parentheses) of the replies given. The symbols ** and * in the table signify a significant difference in t-test results at a significance level of 1% and 5%, respectively.

The results of this experiment show that the average item information of the proposed method was significantly higher than that of time-constrained CAT, Namely, the accuracy of estimating ability was significantly higher by the proposed method than by time-constrained CAT. The results of the questionnaires, meanwhile, revealed that neither travel time nor the number of moves was excessive. The above experimental results demonstrate the effectiveness of the proposed method.

VII. CONCLUSION

We described the development of a mobile testing system using CAT to efficiently assess learning in the field at sites like museums and parks. Since learning assessment in the field requires that the examinee move from one specific position to another, conventional forms of CAT can result in unnecessary back-and-forth movements that simply waste time. Moreover, as the time allotted to a test is generally limited, the test may end before the examinee has had a chance to respond to a sufficient number of test items thereby decreasing the accuracy of estimating ability. To deal with these issues, we proposed a system that makes use of an optimization problem called the traveling purchaser problem (TPP) to optimize travel time in CAT. CAT that uses TPP with a time-limit constraint in this way has the advantage of raising the efficiency of testing that includes movement in the field and improving the accuracy of estimating the examinee's ability. The results of a experiment confirmed the validity of the proposed method for CAT in the field.

REFERENCES

[1] W. V. O. Quine, *Word and Object (Studies in Communication)*, first edition ed. The MIT Press, 3 1964.

[2] L. Wittgenstein, *Philosophical Investigations*, 4th ed. Wiley-Blackwell (an imprint of, 2009.

[3] P. R. Santos, M. Pérez-Sanagustín, D. Hernández-Leo, and J. Blat, "Questinsitu: From tests to routes for assessment in situ activities," *Computers & Education*, vol. 57, no. 4, pp. 2517 – 2534, 2011.

[4] S.-H. Huang, T.-T. Wu, H.-C. Chu, and G.-J. Hwang, "A decision tree approach to conducting dynamic assessment in a context-aware ubiquitous learning environment," in *Wireless, Mobile, and Ubiquitous Technology in Education, 2008. WMUTE 2008. Fifth IEEE International Conference on*, march 2008, pp. 89 –94.

[5] H.-C. Chu, G.-J. Hwang, C.-C. Tsai, and J. C. Tseng, "A two-tier test approach to developing location-aware mobile learning systems for natural science courses," *Computers & Education*, vol. 55, no. 4, pp. 1618 – 1627, 2010.

[6] C.-B. Yao, "Adaptive context aware and intelligent searching in mobile learning applications," in *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, vol. 5, feb. 2010, pp. 802 –806.

[7] C.-M. Chen and C.-J. Chung, "Personalized mobile english vocabulary learning system based on item response theory and learning memory cycle," *Computers & Education*, vol. 51, no. 2, pp. 624 – 645, 2008.

[8] G.-J. Hwang and H.-F. Chang, "A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students," *Computers & Education*, vol. 56, no. 4, pp. 1023 – 1031, 2011.

[9] C.-h. Chen, "The implementation and evaluation of a mobile self- and peer-assessment system," *Comput. Educ.*, vol. 55, no. 1, pp. 229–236, Aug. 2010.

[10] P. Zhang, G. Wills, and L. Gilbert, "Ims qtiengine on android to support mobile learning and assessment," *2010 International Computer Assisted Assessment (CAA) Conference*, 2010.

[11] W. J. van der Linden and C. A. Glas. (2000, 7) Computerized adaptive testing: Theory and practice.

[12] C. N. Mills, M. T. Potenza, J. J. Fremer, and W. C. Ward, Eds., *Computer-Based Testing: Building the Foundation for Future Assessments*. Routledge, 4 2002.

[13] S. A. McIntire and L. A. Miller, *Foundations of Psychological Testing: A Practical Approach*, 2nd ed. Sage Publications, Inc, 11 2006.

[14] D. N. M. de Gruijter and L. J. T. van der Kamp, *Statistical Test Theory for the Behavioral Sciences (Chapman & HallCRC Statistics in the Social and Behavioral Sciences)*, 1st ed. Chapman and HallCRC, 8 2007.

[15] T. J. Kline, *Psychological Testing: A Practical Approach to Design and Evaluation*, 1st ed. Sage Publications, Inc, 2 2005.

[16] W. J. van der Linden, D. J. Scrams, and D. L. Schnipke, "Using response-time constraints to control for differential speededness in computerized adaptive testing," *Applied Psychological Measurement*, vol. 23, no. 3, pp. 195–210, 1999.

[17] W. J. van der Linden, *Linear Models for Optimal Test Design (Statistics for Social and Behavioral Sciences) [Hardcover]*. Springer, 2005.

[18] E. Triantafillou, E. Georgiadou, and A. A. Economides, "The design and evaluation of a computerized adaptive test on mobile devices," *Computers & Education*, vol. 50, no. 4, pp. 1319 – 1330, 2008.

[19] Y.-M. Huang, Y.-T. Lin, and S.-C. Cheng, "An adaptive testing system for supporting versatile educational assessment," *Computers & Education*, vol. 52, no. 1, pp. 53 – 67, 2009.

[20] T. Ramesh, "Travelling purchaser problem," *Opsearch*, vol. 18, pp. 78–91, June 1981.

[21] Y. MIYASAWA and M. UENO, "Mobile testing for authentic assessment in the field," *Artificial Intelligence in Education (AIED 2013)*, 2013.