USE OF COMPUTERIZED ADAPTIVE TESTING IN CORPORATE E-LEARNING

Kemal Kačapor[1], Dušanka Milanov [2]
School of Economics and Business, Sarajevo, University of Sarajevo, Bosnia and Herzegovina[1]
Faculty of Sciences, Novi Sad, University of Novi Sad, Serbia[2]
kemal.kacapor@efsa.unsa.ba[1], dusanka.milanov@gmail.com[2]

Abstract *- This paper will describe the potential of computerized adaptive testing, as a part of E-learning, within a corporate context. The success of any organization today is more and more dependent on the quality of the people they are able to attract, recruit and retain and employee value is essential for competitive market environment. Adaptive tests are designed to maximize measurement efficiency, or the precision of test scores in relation to test length. Therefore, the potential of computerized adaptive testing in determining employees knowledge and skills development, which are vital in any organization, is going to be presented and analyzed.*

1. INTRODUCTION

Employment and training tests in corporate environment are mostly administered by computer but using a traditional fixed testing format in which all examinees are administered the same items. In corporate learning, where time is essential, assessments have to be flexible and precise and take every advantage of modern technologies. Computers as an administration devices can support this goal by intelligently administering and scoring items using Computerized Adaptive Testing (CAT).

2. COMPUTERIZED ADAPTIVE TESTING

In the classical "paper and pencil" format, tests or assessments are identical for all the examinees and usually have a predetermined duration and a fixed number of questions which have various levels of difficulty. The examinee's mark is usually the sum of the scores obtained in each of the questions and is used as competence criterion. Science recognized long time ago that giving a test that is much too easy for the candidates can turn out to be a waste of time, and can lead to unwanted candidate behavior such as careless mistakes or deliberately choosing incorrect answers that might be the answers to "trick questions". On the other hand, questions that are much too hard, can produce generally inaccurate test results, because candidates try to attempt to answer the questions, often by guessing. Finally, the mark doesn't give enough information about the student answers which could help to determine unusual responses patterns or gaming behavior from the learner; features that are important especially for diagnostic testing. Despite these problems, in most of the E-learning platforms, the tests often have the traditional testing format and are implemented as a simple online "paper and pencil" version.

This issue is addresses with use of Computer adaptive testing (CAT) which is a special case of Computer based testing (CBT), based on interactive method for assessing the level of a test-taker's knowledge, proficiency, ability, or performance using questions tailored to the specific student. The CAT system selects questions from a pool of recalibrated items appropriate for the level of the specific student.

A CAT system tailors the test to the proficiency of the individual examinee meaning it adjusts the test by presenting easy questions to a low-proficiency examinee and difficult questions to a high-proficiency examinee. As an alternative to giving each examinee the same fixed test, CAT item selection adapts to the ability level of individual examinees.

After each response the ability estimate is updated and the next item is selected such that it has optimal properties according to the new estimate [3]. The CAT first presents an item of moderate difficulty in order to initially assess each individual's level. During the test, each answer is scored immediately and if the examinee answers correctly, then the test statistically estimates examinee's ability as higher and then presents an item that matches this higher ability. If the next item is again answered correctly, it re-estimates the ability as higher still and presents the next item to match the new ability estimate. The opposite occurs if the item is answered incorrectly. The computer continuously re-evaluates the ability of the examinee until the accuracy of the estimate reaches a statistically acceptable level or when some limit is reached; such as a maximum number of test items.

The overall score is determined from the level of the difficulty, and as a result, while all examinees may answer the same percentage of questions correctly, the high-ability ones will get a better score as they correctly answer more difficult items. Overall score of each examinee depends not only on the percentage of questions answered correctly but also on the difficulty level of these questions. Even if both examinees answer the same percentage of questions correctly, the high-proficiency examinee gets a higher score because the examinee answers correctly more difficult questions. Because each test is tailored to the individual examinee, far more information is gained from the examinee's response to each item than in conventional test.

A. Item Response Theory

CAT require defining computational models which allow objective estimation and comparison of the proficiencies

of learners that received different questions during a test. The psychometric technology that allows equitable scores to be computed across different sets of items is item response theory (IRT). IRT is also the preferred methodology for selecting optimal items which are typically selected on the basis of information rather than difficulty.

IRT is a statistical framework in which examinees can be described by a set of one or more ability scores that are predictive, through mathematical models, linking actual performance on test items, item statistics, and examinee abilities. It is a set of related psychometric models that provides a foundation for scaling persons and items based on responses to assessment items [6]. The person parameter usually is the proficiency or cognitive ability within a specific domain and it is represented by the Greek letter θ. A question in a test is a simple example of an item.

Much of the literature on IRT focuses on its models. Those models are usually functions relating person parameter and item parameters to the probability of a discrete outcome, such as a correct response to that item. The three parameters logistic (3PL) model, first described by Birnbaum (1968), is the most widely used because it gives enough flexibility for implementation. Under the 3 parameter IRT model, the probability of a correct response to a given item i is a function of an examinee's true ability and three item parameters. Those three item parameters are the discrimination a, the difficulty b and the pseudo-guessing c. The later represents the chance for a low level examinee to find by guessing the correct response to the item. The conditional probability for a person with an ability θ to get a correct response to an item i (ai, bi and ci) is given by the equation bellow.

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}}$$

IRT provides strategies for estimating:

a. The *ai*, or item discrimination, parameter,
b. The *bi*, or item difficulty, parameter, and
c. The *ci*, or guessing, parameter.

IRT has been successfully used in some computer adaptive testing system [1]. IRT is a data oriented model and different studies show its advantages in term of implementation complexity, predictive power and independence from a subjective expert appreciation.

B. CAT Algorhytm

The CAT procedure is very simple and can be explained as: a test-taker is estimated (or guessed) to have a certain ability. An item of the equivalent level of difficulty is asked. If the test-taker succeeds on the item, the ability estimate is raised. If the test-taker fails in the item, the ability estimate is lowered. Another item is asked, targeted on the revised ability estimate and the process repeats.

Computer adaptive testing can begin when an item bank (or question pool) exists with IRT item statistics available on all items, when a procedure has been selected for obtaining ability estimates based upon candidate item performance, and when there is an algorithm chosen for sequencing the set of test items to be administered to candidates.

The CAT algorithm is usually an iterative process with the following steps:

1. All the items that have not yet been administered are evaluated to determine which will be the best one to administer next given the currently estimated ability level.
2. The "best" next item is administered and the examinee responds.
3. A new ability estimate is computed based on the responses to all of the administered items.
4. Steps 1 through 3 are repeated until a stopping criterion is met.

There are five technical components in building a CAT [7]:

1. Calibrated item pool - A pool of items (questions) must be available for the CAT to choose from. The pool must be calibrated with a psychometric model, usually IRT model.
2. Starting point or entry level - If some previous information regarding the examinee is known, it can be used, but often the CAT just assumes that the examinee is of average ability - hence the first item often being of medium difficulty.
3. Item selection algorithm - If the CAT has an estimate of examinee ability, it is able to select an item that is most appropriate for that estimate which is done by selecting the item with the greatest information at that point.
4. Scoring procedure - After an item is administered, the CAT updates its estimate of the examinee's ability level. If the examinee answered the item correctly, the CAT will likely estimate their ability to be somewhat higher, and vice versa. This is done by using the item response function from item response theory to obtain a likelihood function of the examinee's ability.
5. Termination criterion - The decision as to when to stop a CAT test is the most crucial element. If the test is too short, then the ability estimate may be inaccurate. If the test is too long, then time and resources are wasted, and the items exposed unnecessarily. The test-taker also may tire, and drop in performance level, leading to invalid test results.

The CAT test stops when:

1. The item bank is exhausted.
2. The maximum test length is reached.
3. The ability measure is estimated with sufficient precision.
4. The ability measure is far enough away from the pass-fail criterion.
5. The test-taker is exhibiting off-test behaviour.

The CAT test cannot stop before:

1. A minimum number of items has been given.
2. Every test topic area has been covered.
3. Sufficient items have been administered to maintain test validity under challenge or review.

3. USE OF CAT IN CORPORATE ENVIRONMENT

Most organizations have numerous challenges when hiring new candidates, conducting job skills assessments as well as estimating future skills needs. Questions which they face during these processes are about improving the quality of candidates employed; areas of training which are really needed across the enterprise; identifying the individuals who need training within the organization; ensuring that training programs are indeed effective and more.

Incorporating a CAT based assessment solution into daily recruitment process, skill audits, project staffing and training can enable companies to unleash the full potential of employee base and contribute to competitive advantage. Today's solutions offer assessments which allow conducting employment screening using pre-hire testing; identifying specific training requirements with pre-training assessments; evaluating training effectiveness and knowledge transfer with post-training assessments; providing career development with employment testing and certification programs etc.

Placing the person with the right skills and knowledge in the right position can significantly increase productivity and reduce staff turnover. Human resource experts estimate the costs associated with correcting hiring errors range between 1 to 2 times the candidate's annual salary. Using a knowledge measurement solution to establish candidate suitability, organizations can reduce the administration time associated with hiring, increase the efficiency of the hiring process itself and improve the quality of candidate selection.

In corporate education, we can differentiate between adaptive testing that is used for certification, to assess individual achievement and to rank candidates, and adaptive testing that is used to assist employee learning/training, which focuses more on customizing the testing to inform learning and is part of the developmental process. This can be used by both the learner and the teacher: learners can observe their personal learning progress and decide how further to direct their learning process and tutors can individually support learners and formulate judgments about the quality and effectiveness of the provided instruction. In this case adaptive testing can provide important diagnostic tools that contribute meaningfully to learning.

In a corporate setting, the principles of adaptive learning can be applied to the entire organization and not just the individual learner. To the business unit, adaptive learning is largely oriented towards contributing lessons learned for the greater common goal. To the organization, adaptive learning is considered a vital, strategic component which can be used for "learning from mistakes", with a goal of establishing appropriate mechanisms to preserve and nurture the evolving corporate growth [5].

Most organizations execute generic training programs for entire departments without truly understanding what areas of training are required, and by whom. After much time and investment in training program execution, there are typically no mechanisms in place to determine overall training effectiveness and return on investment (ROI). Using a knowledge measurement solution to establish and evaluate training programs, companies could be able to determine who in the organization needs training; understand what areas of training should focus on; optimize investment balance between training and organizational skill needs etc.

4. ADVANTAGES AND BARRIERS OF CAT USAGE

In terms of corporate benefits, CAT can help achieve the following: increase employee competence (improve employee productivity, hiring and selection process, reduce employee attrition, increase customer credibility); increase job satisfaction (reduce training time and cost, improve training ROI, provide objective ROI confirmation); reduce hiring time and costs (reduce staff attrition, use objective benchmarks to differentiate candidates).

Also, no time is wasted on questions not suitable for the assessment-taker; assessment-takers continually challenged by questions at their demonstrated knowledge level; and quicker assessments are less disruptive to employer, employee and candidates.

In general, computerized testing greatly increases the flexibility of test management. Some of the benefits are:

- Tests are given "on demand" and scores are available immediately.
- Neither answer sheets nor trained test administrators are needed. Test administrator differences are eliminated as a factor in measurement error.
- Tests are individually paced so that a examinee does not have to wait for others to finish before going on to the next section. Self-paced administration also offers extra time for examinees who need it, potentially reducing one source of test anxiety.
- Test security may be increased because hard copy test booklets are never compromised.
- Computerized testing offers a number of options for timing and formatting. Therefore it has the potential to accommodate a wider range of item types.
- Significantly less time is needed to administer CATs than fixed-item tests since fewer items are needed to achieve acceptable accuracy. CATs can reduce testing time by more than 50% while maintaining the same level of reliability. Shorter testing times also reduce fatigue, a factor that can significantly affect an examinee's test results.
- CATs can provide accurate scores over a wide range of abilities while traditional tests are usually most accurate for average examinees.

The first barrier encountered in CAT is the calibration of the item pool [6]. In order to model the characteristics of the items (e.g., to pick the optimal item), all the items of the test must be pre-administered to a sizable sample and then analyzed. To achieve this, new items must be mixed into the operational items of an exam (the responses are recorded but do not contribute to the test-takers' scores), called "pilot testing" or "pre-testing". This presents logistical, ethical, and security issues. For example, it is impossible to field an operational adaptive test with brand-new, unseen items; all items must be pretested with a large enough sample to obtain stable item statistics.

Some other CAT limitations also raise several technical and procedural issues:

- CATs are not applicable for all subjects and skills. Most CATs are based on an item-response theory model, yet item response theory is not applicable to all skills and item types.
- Hardware limitations may restrict the types of items that can be administered by computer. Items involving detailed art work and graphs or extensive reading passages, for example, may be hard to present.
- CATs require careful item calibration. The item parameters used in a paper and pencil testing may not hold with a computer adaptive test.
- CATs are only manageable if a facility has enough computers for a large number of examinees and the examinees are at least partially computer-literate. This can be a big limitation.
- The test administration procedures are different. This may cause problems for some examinees.
- With each examinee receiving a different set of questions, there can be perceived inequities.
- Examinees are not usually permitted to go back and change answers.

5. CONCLUSION

In this time of global market competition, rapid technological advances, demographic changes, and service/knowledge-based economy force, organizations need to educate and constantly train their workforce.

Although it has taken some time for the business world to understand the benefits of E-learning, it is clear that limitless opportunities are ahead. Employing a successful e-learning strategy allows a corporation to cut costs significantly, while increasing workplace satisfaction and raising employee motivation.

Adaptive tests are designed to maximize measurement efficiency, or the precision of test scores in relation to test length. This means an adaptive test can either save time by being shorter than a conventional test of equal precision, or improve score quality by being more precise than a conventional test of equal length. The examinees with the most to gain are those at either the high or low extremes of the performance continuum. They are usually poorly served by conventional tests, which are generally designed to best fit the average test-taker.

ACKNOWLEDGMENT

This work is financially supported by Ministry of Science and Technological Development, Republic of Serbia; under the project number TR32044 “The development of software tools for business process analysis and improvement”, 2011-2014.

REFERENCE

[1] Baker, F. “The Basics of Item Response Theory”, ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, 2001.

[2] Economides, A., Roupas C. “Evaluation of Computer Adaptive Testing Systems”, International Journal of Web Web-Based Learning and Teaching Technologies, Vol. 2, Issue 1, pp. 70-87, 2007.

[3] Linden, J. Glass, C. “Computerized Adaptive Testing: Theory and Practice”, Kluwer Academic Publishers, Netherlands, 2003.

[4] Liu, L., LaMont Johnson, D., Maddux, C., Henderson, N. “Evaluation and Assessment in Educational Information Technology”, The Haworth Press, Inc., 2001.

[5] Makransky, G. “Computerized Adaptive Testing in Industrial and Organizational Psychology”, University of Twente, 2012.

[6] Wainer, H. “Computerized Adaptive Testing: A primer”, Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

[7] Weiss, D., Kingsbury, C. “Application of Computerized Adaptive Testing to Educational Problems”, Journal of Educational Measurement, Vol. 21, Issue 4, pp. 361-375, 1984.