COMPUTER-ADAPTIVE TESTING IN SCIENCE EDUCATION

Elena C. Papanastasiou

ABSTRACT
Computer Based Learning (CBL) can have great potential if used appropriately to enhance the quality of science education. However, this quality can be enhanced even further with the use of Computer Based Testing (CBT), and more specifically, with Computer Adaptive Testing (CAT). For the purpose of this paper, the author will describe how computer adaptive testing works, as well as its advantages in relationship to how it can improve the CBL process in the subject area of science. Some limitations and issues that educators need to take into consideration before implementing a science education CAT will also be discussed.

KEYWORDS
Computer adaptive testing, CAT, computer based testing, computer based learning, science education, evaluation, formative feedback

INTRODUCTION

Computer based learning has the extreme potential to improve instruction in numerous subject areas and disciplines, including the subject area of science. However, computer based learning needs to be closely and constantly monitored to ensure its effectiveness. This is especially the case since some prior studies have found computer use to be negatively correlated with achievement in mathematics and science (Papanastasiou, 2002b; Papanastasiou & Ferdig, 2003). Although it is not clear under what circumstances this negative relationship has developed, and if there is a causal relationship between these variables, it still exists. Consequently, this relationship should remind educators that computer use is not necessarily a 'panacea', and that it should not be used irresponsibly to occupy the attention of students who are hard to deal with.

This negative relationship between computer use and achievement should also remind educators of the significant need for continuous formative and summative assessment in science, just like in any learning context. By using assessment properly, problems that may arise from any learning situation can be identified and possibly corrected if they are noticed early enough. However, assessment also needs to be used thoughtfully in a way that it can complement the learning process. Since computer based learning is a focus of this conference, it will be tied with computer based assessment for this paper. So the purpose of this paper is to go beyond mere computer based learning, to describe computer adaptive testing, and discuss its effects, its advantages, and how it can effectively complement computer based learning in the subject area of science.

Definition

Computer based testing (CBT) could be defined as any type of assessment that is administered through the computer. However, computer based testing can encompass many forms, depending on how adaptable the test is on the item level (The College Board, 2000). For example, some CBT, which are also called computerized fixed tests, are purely linear (Parshall, Spray, Kalohn & Davey, 2002). These are the tests that most closely resemble paper and pencil tests, since they are fixed form, fixed length, and the test items are organized in advance and placed in a predetermined order. In contrast to

computerized fixed tests, computer adaptive tests (CAT) are the computer based tests that have the maximum degree of adaptivity since they can be adapted for each examinee, based on the amount, difficulty and order in which the items are administered to each examinee. So computer adaptive tests (CAT) could be defined as the computer based tests which are created and adapted specifically for each examinee based on the examinee's ability estimate, and based on the way in which each examinee has responded to the previous items that have been administered to them.

Computer adaptive testing is based on the theories and advances of Item Response Theory (IRT). More specifically, in contrast to classical test theory, IRT has managed to put the examinee ability, and item characteristics on the same continuum. Consequently, item characteristic curves (ICCs) in IRT display the probability of answering an item correctly, according to any location that a person can have on the ability continuum. This probability depends on the a-, b-, and c-parameters that are used to characterize each item. These parameters are obtained from the pilot testing of the items on a sample of students before the final version of the test is developed. The a-parameter reflects the rate at which each item can discriminate a successful performance from a non-successful performance, based on the proficiency level of the examinee. The b-parameter reflects the difficulty of each item, while the c-parameter (the pseudo-guessing) parameter reflects the probability of being able to answer an item correctly without having any knowledge on how to answer that item. Equation 1 (Lord, 1980) below is the formula used to determine the probability of answering each item correctly, according to each examinee's ability, and according to the item parameters.

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{\{-1.7a_i(\theta - b_i)\}}} \quad (1)$$

where i represents each individual item and
where θ represents each examinee's true ability which we are trying to estimate.

Once the parameters of each item are estimated from the pilot testing of the items, it is possible to start developing an adaptive test. Typically, computer adaptive tests start by administering an item of about average difficulty to the examinees. This is done because the ability of most examinees lies in the middle of the ability continuum. Consequently, the majority of the students or examinees are able to respond to such average difficulty items, because they will not be too easy and too difficult for them. After the first item is answered, and based on if it is correct or incorrect, the second item is administered. If an examinee A, has responded to the first item on a science computer adaptive test correctly, that indicates to the adaptive test that this examinee has above average ability in science. Consequently, the computer will administer a second, more difficult item to that examinee. If this examinee also answers the second item correctly, then the third item administered will be even more difficult that the first two items. The goal is to be able to administer a test whose difficulty matches the achievement level of the examinee. When this is achieved, the ability of the examinee can be estimated with small margins of error.

If another examinee B had responded incorrectly to the first item, that indicates to the adaptive test that the science ability of this student is below average. So in an attempt to find out where the ability level of this examinee really is, the next item that will be administered will be easier than the first one. If examinee B also responds incorrectly to the second item, the third item that will be administered will be even easier. This process will continue to take place until an accurate estimate of the examinee's science ability is obtained.

This maximum amount of adaptivity which is associated with computer adaptive tests, results in the many advantages that are associated with CATs. Consequently, this paper will focus specifically on computer *adaptive* testing, which is hypothesized to be the most advantageous and efficient for complementing the computer based learning process in the subject area of science.

ADVANTAGES OF COMPUTER ADAPTIVE TESTS

The main advantages of computer adaptive testing stem from the fact that they are efficient in terms of items used, time, as well as resources. These advantages will be discussed briefly in the following section from the perspective of the examinees or students who take such tests, from the perspective of the educator who wants to identify what each student really knows, as well as from the perspective of a test developer who develops such tests.

Item efficiency

Computer adaptive tests have the capability of estimating an examinee's ability more accurately and with fewer items than with paper and pencil tests. Typical paper and pencil tests are created for mass testing, so that the same test can be used by a large group of students who all vary in ability and achievement level. In order to do this, the majority of items on these tests include a large proportion of average difficulty items (since the majority of the students are in the middle of the ability continuum) while a few low and some high ability items are also administered (since there are fewer students in the two extremes of the ability continuum). As a result, this type of test content creates problems to the high and low ability examinees. When a low ability examinee is responding to such fixed-length tests, that student might be able to respond to the first few items which are relatively easy. As the test progresses however, the average and high difficulty items might be too difficult for that examinee. Consequently, that student might end up guessing the answers to those items, or might just leave them blank. In this scenario, it would be too difficult to reach valid and reliable conclusions about what this student might know on this science test, since any conclusions would have to be based only on the first few items that the student was able to complete. So overall, trying to make any inferences about a student's ability from only a couple of questions would be quite difficult.

Another more specific example of the same scenario is presented below. A school district wants to create a biology test that focuses on the parts and functions of the liver. On this test, the low difficulty questions have to do with locating the liver on pictures of a human body, while the most difficult questions have to do with trying to diagnose what might be wrong with the liver based on a different set of pictures. In this case, if a student cannot even identify or locate the liver on a picture of the human body, there would be no reason to administer the more difficult questions to that student since it would be clear that the student would not be able to predict what problems a liver has by looking at those sets of pictures of a malfunctioning liver.

When examining such tests from the perspective of a high achieving examinee, the situation is slightly better, although it's still not perfect. When a high achieving examinee takes linear test, in most cases the majority of the items will be too easy for this person. For example, let's take the case of a science placement test that tries to place students in levels of a chemistry class. This test could contain items that range from asking the students to identify the symbols of elements on the periodical table, to items where the students have to solve chemistry equations. If we were interested in identifying the ability level of a high achieving student to determine if they should enter the proficient or advanced chemistry class, the ideal situation would be to administer many items that are related to the content areas of those two classes (e.g. by solving chemistry equations). However, there would be no reason to ask such a student questions about what is the symbol that represents the element of Carbon since that would purely be a waste of time and effort. In addition, by administering items that are too easy for such a student, extraneous variables could be entered into their ability estimate, such as careless errors induced by boredom (Wainer, 2000).

Consequently, instead of administering a single test to all students with questions that are too easy or too difficult, adaptive tests efficiently allow the administration of questions that are specifically targeted at each examinee's ability level. When all of the items that are administered to are exactly targeted at each student's ability level, an educator or test administrator can reach more reliable and valid conclusions about the actual knowledge that each student possesses. In addition, by administering items

that are specifically targeted at an examinee's ability, the examinee's final score can be estimated more accurately in less time, by administering fewer items on a CAT rather than on a paper and pencil test.

Feedback

Another one of the advantages of computer based tests in general, as well as of computer adaptive tests, is that they can administer direct and immediate feedback to the student/examinee as well as to the teacher (Wise & Plake, 1990). With typical paper and pencil tests, there always tends to be a lag of time between the administration of the test and of its scoring. As a result, most of these tests ended up being used only summatively in an attempt to assign grades to the students. Educationally this is not a very sound practice since it does not allow the use of formative feedback that could supplement the learning process. Without this formative assessment process, educators would not be able to determine if computer based learning is actually helping all students learn in science, or not. This is especially worrisome because without proper assessment, some students could be disadvantaged from computer based learning. If these cases are not identified early enough, the negative consequences that it could have on the students could exponentially increase as time goes by.

In addition to the total score that such assessments can provide, which indicates how each student has performed overall, they can also provide a list of the content areas and objectives that have been met by each student, based on their performance on the adaptive test. If certain content area objectives have not been met by each student, that assessment can in turn be immediately linked back to specific units of the computer based learning process so that the students can meet those objectives as well.

However, one issue that could arise with continuous testing for the teacher, is the possibility that some students could memorize test items, and inform other students about them. However, if the adaptive test has a relatively large item pool, the memorization of test items should not be a problem especially in light of the fact that different students should get different test items based on their individual ability level.

Time

Although CATs are more difficult and time consuming to create, they are more time efficient from the perspective of the test-taker, and of the educator or test administrator. More specifically, from the perspective of the examinee, students have to respond to fewer questions on adaptive tests than with regular paper and pencil tests. Therefore, such tests might not tire them as much since they will have to respond to fewer questions that are specifically targeted to their ability level. Moreover, paper and pencil tests are typically administered through a whole group administration, at a time which could have been inconvenient for some students. During that whole group administration of a test, all students have to wait until all students have completed the test before they can start any other activity. With CATs however, the students can take the test whenever they are able to do so, as long as a computer is available, and they do not have to wait until the whole group or class is ready to take the test, or until the whole group has finished taking the test.

From the perspective of the teacher, adaptive testing is efficient in terms of time as well. First of all, teachers will no longer have to worry about creating tests for their classroom as long as the CAT that they use matches the science content areas that they teach. In addition, teachers can also save time from grading tests since the students' answers on CATs are saved and scored electronically. This is the case for selection type items (e.g. multiple choice, true-false) as well as supply type items such as short answer or essay type questions (Bennett, 1999) that are much harder to grade by the teacher.

More flexibility in item types

The use of the computer for testing in general, has the advantage of being able to use more flexible and creative item types than regular paper and pencil tests (Parshall, Stewart & Ritter, 1996). Tests today no longer have to be confined to pure text items or to items that might include a few pictures. Test items for computer based tests can include high resolution pictures, movies with motion and sound, as well as voice synthesizers, and oral comprehension of spoken language (Parshall, Spray, Kalohn & Davey,

2002; Wainer, 2000). This is especially useful when combined with computer based learning because adaptive tests could also include test items obtained from examples or experiments used during the computer based learning process. These types of items, when combined with their multimedia properties could make the testing situation a bit more realistic and more similar to the actual learning process. (This is especially useful for the subject area of science, since the computer could even replace the science lab when science software can be used to perform experiments which are too dangerous or expensive to perform otherwise).

In addition to making the testing situation more realistic, such items also have the advantage of eliciting positive attitudes from the students who take such tests on the computer (Parshall, Stewart & Ritter, 1996). Consequently, the students end up being more motivated to do their best, which in turn increases the reliability of the student's test scores.

Accommodations for students with special needs

Administering tests on the computer has the advantage of being able to accommodate students with various forms of special needs (O'Neill, 1995). This is not a special function of CATs, but a function of computer based tests in general. More specifically, because of the advanced multimedia properties of technology today, computers have the capability of easily enabling us to teach, as well as test students who could have various types of disabilities such as visual or hearing impairments. For example, CATs could easily assign more time for each question to students with disabilities, as well as use sound, motion or text to assist students appropriately, based on their impairment. These types of accommodations can alleviate some of the frustration that is typically faced by such students when they are integrated in regular classrooms, which in turn could assist them in demonstrating their actual knowledge more accurately through computer adaptive tests.

Other advantages

Computer adaptive tests also include some additional advantages that will not be discussed in length. These include the fact that CATs enhance test security, since booklets can no longer be stolen, and since trying to copy from people nearby would not be beneficial since most of the items on the tests are different for each examinee. In addition, through CATs, other types of data can be collected such as the amount of time needed to answer each question, and the number of changes that the students made to their answers on the test, if that is permitted. However, these issues are besides the scope of this paper and will not be discussed further in more detail.

ISSUES FOR CONSIDERATION

Besides the advantages of computer adaptive tests, there are also some issues associated with these tests that should be taken into consideration and should not be ignored before the decision to create or use such tests are adopted. These are the issues of computer familiarity by the students, the issue of cost for the teacher or school who would adopt such an assessment, as well as the issue of the assessment and curriculum content match.

Student familiarity with computer use

An important issue that should be taken into account before implementing a CAT in a school setting is the amount of familiarity and comfort that students have with the use of computers. There are students who might have limited access to computers at home, or who might even have anxiety in relation to computer use. In these cases, such students will not be able to perform optimally either through the computer based learning process, or through a CAT. More specifically, it is likely that students who are not familiar with computers or who have computer anxiety might focus more on the technology that is in front of them and on how to work with it rather than with the actual subject matter activities (e.g. science learning or science assessment). Consequently, their learning, or their estimate of their science abilities could be masked because of their computer anxiety or unfamiliarity. This anxiety and unfamiliarity could add additional sources of error to the examinee's ability estimate, which could further bias any conclusions reached for these examinees in relation to their science knowledge.

However, when computer adaptive testing naturally follows computer based learning, such errors tend to be minimized.

Cost

The initial cost of setting up a CAT can be significant (Meijer & Nerling, 1999). This entails the cost for the purchase of the computer software, hardware, as well as the costs for the creation, set up, and maintenance of a valid and reliable CAT (The College Board, 2000). However, if the computers are already in place in an educational setting because computer based learning is already used, then a large part of the initial CAT setup cost can been eliminated. In addition, it is also likely that CATs can reduce some of the costs associated with paper and pencil tests (Wise & Plake, 1990). For example, the cost for printing and scoring paper and pencil tests can be eliminated this way, since these are done automatically through a CAT. So although the costs in relation to using a CAT cannot be eliminated completely, it can be slightly reduced in the long run.

Assessment and curriculum content match

Another main issue that needs to be taken into account is the content match of the CAT with the school science curriculum, as well as with the taught science curriculum. With regular teacher made tests, the content match between what was tested and what was taught was always there. (These tests might have had other psychometric problems with their reliability and validity, but at least the content match was always there). However, CATs are more difficult and more time consuming to create. Consequently, educators or school districts would have to rely on pre-made computer adaptive tests. In that case, there comes the issue of assessment driven instruction. Would it be acceptable to base the materials that will be taught in science on the content that is covered on the computer adaptive test? This is an issue that is besides the scope of this paper. However, any decisions that will be made on this topic should always be backed up by research evidence to ensure that any scores obtained by such tests are reliable and valid within the educational setting that they are used.

Other unresolved issues

Since adaptive testing is a relatively young field, there are a lot of issues related to its applications that have not been conclusively resolved yet. For example, many adaptive tests do not allow examinees to omit items or change any of their answers throughout the test (Papanastasiou, 2002a). So if a student has made a careless error on an item, and later on realized that mistake, the CAT might not allow that student to change that answer. In addition, students that have been used to skipping items that appeared too difficult so that they can tackle them after they reached the end of the test, cannot do so anymore. This is because omitting items, and item review does not follow the logic on which adaptive tests are based on. In addition, if item review were allowed, it would be possible for the students to use cheating strategies that already exist in the measurement literature in order to increase their test scores (Kingsbury, 1996, Wainer, 1993). However, if adaptive testing is used for low stakes assessment within the context of a science class, such cheating attempts would be less likely to occur. If it is expected that such cheating attempts will be less likely to occur, educators could try to modify the test interface in order to allow such response options.

CONCLUSION

Current research in the field of measurement and testing have shown the increased potential of computer adaptive tests. In addition, the recent trends in computer based learning, and the integration of technology in many schools have created a boost in computer based testing, which will only increase further in the future (The College Board, 2000). However, computer adaptive testing as well as its advantages and possibilities go a step beyond that. This can be seen from the ever increasing number of large scale tests (e.g. GRE, TOEFL, ASVAB) that have become or are becoming adaptive (Papanastasiou, 2001). What needs to be done now is for educators, and especially science educators to start taking advantage of these possibilities, by using computer adaptive testing to complement their teaching, in combination with the use of computer based learning in their science classes. However,

such steps always need to be taken slowly and wisely to ensure that the assessment procedures are well integrated with the computer based learning process to ensure its maximum effectiveness.

REFERENCES

Bennett, R. E. (1999). Using new technology to improve assessment. RR99-6. Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum associates.

Meijer, R. R. & Nerling, M. L. (1999). Computerized adaptive testing: Overview and introduction. Applied psychological measurement, 23(3), 187-194.

O'Neill, K. (1995). Performance of examinees with disabilities on computer-based academic skills tests. Paper presented at the American educational research association, San Francisco, April, 1995.

Papanastasiou, E. C. (2001). A 'Rearrangement Procedure' for administering adaptive tests when review options are permitted. (Doctoral dissertation, Michigan State University, 2001).

Papanastasiou, E. (2002a). A 'rearrangement procedure' for scoring adaptive tests with review options. Paper presented at the National Council of Measurement in Education, New Orleans, LA.

Papanastasiou, E. (2002b). Factors that differentiate mathematics students in Cyprus, Hong Kong, and the USA. Educational Research and Evaluation, 8 (1), 129-146.

Papanastasiou, E. C. & Ferdig, R. E. (2003, January). Computer use and mathematical literacy. An analysis of existing and potential relationships. Paper presented at the third Mediterranean conference on mathematics education, Athens, Greece, January 3-5, 2003.

Parshall, C. G., Spray, J. A., Kalohn, J. C. & Davey, T. (2002). Practical considerations in computer-based testing. NY: Springer.

Parshall, C. G., Stewart, R. & Ritter, J. (1996). Innovations: Graphics, sound and alternative response modes. Paper presented at the National Council of Measurement in Education, April 9-11, 1996, New York.

The College Board. (2000, April). An overview of computer-based testing. RN-09.

Wainer, H. (2000). CATs: Whither and whence. Psicologica, 21(1-2), 121-133.

Wise, S. L. & Plake, B. S. (1990). Computer-based testing in higher education. Measurement and evaluation in counseling and development, 23, 3-10.

Elena C. Papanastasiou, Ph.D.
University of Kansas and University of Cyprus
Department of Education
P.O. Box 20537
1678 Nicosia
Cyprus
Email: elagatha@cytanet.com.cy