Title: Validity of computerised adaptive tests for biology achievement testing
Author(s): Chew Lee Chin
Source: ERA Conference, Singapore, 24-26 November 1997
Organised by: Educational Research Association of Singapore (ERAS)

This document may be used for private study or research purpose only. This document or any part of it may not be duplicated and/or distributed without permission of the copyright owner.

The Singapore Copyright Act applies to the use of this document.
Validity of Computerised Adaptive Tests For Biology Achievement Testing

Chew Lee Chin
National Institute of Education
Nanyang Technological University

Introduction

The recent item response theory (IRT) research and advances in computer technology have led to computerised adaptive testing (CAT). A growing body of research has indicated that computerised adaptive tests (CATs) can be used to obtain a precise and accurate estimate of an examinee's ability on an underlying measurement scale (Bugbee, 1996; Kingsbury & Houser, 1993; Lord, 1977; Weiss, 1982). These customised tests are constructed adaptively during the process of test administration with test items drawn from a previously IRT-calibrated item bank. That is, the computer selects and administers each new test item by first estimating a level of ability based on the examinee's response to a previous test item.

In the United States, bolstered by research showing that CAT has several advantages over that of conventional paper-and-pencil testing, large-scale testing of the Graduate Record Examination (Vogel, 1994) as well as educational testing at school or college levels (Legg & Buhr, 1992; Shermis et al., 1996) have begun to use CATs for assessing skills proficiency and/or learning outcomes. However, the research community is cognizant of differences between the two testing methods in terms of measurement theory (item response theory versus classical test theory), mode of testing (adaptive versus fixed-item fixed-length tests) and medium of testing (computer-administered versus paper-administered). Because of these significant differences researchers have advocated the need for validation studies to establish the equivalence of computer-based and paper-and-pencil test scores obtained for a particular assessment.

According to the Standards for Educational and Psychological Testing (1985), validity is the “most important consideration in test evaluation” (p. 9). Researchers agree that the evaluation of the test quality should be an important focus of any test. For instance, Popham (1995) contended that validity deals with the defensibility of score-based inferences. The idea is that no test is valid or invalid in itself. Only its use in some application merits a designation of validity. In the case of CAT, Hambleton and Swaminathan (1985) have cautioned that the availability of an ability scale does not ensure its interpretability. Thus, validation studies are needed to determine whether or not the ability scores obtained with CAT serve their intended function.

This present study reports on the accumulated evidence to support the adequacy and appropriateness of the interpretations and use of test scores obtained by CATs for the assessment of biology learning outcomes. The data were collected in conjunction with a CAT implementation to assess students' learning outcomes at the end of a two-year biology course. Evidences of criterion-related validity were gathered to demonstrate that the biology test scores obtained by CAT (CAT_bayes) are systematically related to three other biology achievement scores. These include (a) a paper-and-pencil test that was developed from a same IRT-calibrated item bank (PPT_bayes), (b) and two formal assessments taken by the study sample, namely, the school's final semestral examination (SSE_%total) and the GCE 'O' level examination (GCE_grade). However, it is important to note that both the computerised adaptive test and the paper-and-pencil test used only multiple-choice items while the other two examinations each consisted of written (multiple-choice items, and free-response questions) and practical tests (performance-based assessment items). Nevertheless, these examinations were deemed appropriate sources of validity evidence.
Method

Sample

The 113 participants were secondary students (aged 16+) from a Singapore school who had just completed a 2-year biology course. They were scheduled to take the biology examination at the 1995 GCE ‘O’ levels.

Measures

Two biology achievement tests were used to collect the data. One was a computer adaptive test, and the other a conventional paper-and-pencil test. Both tests were constructed using test items from a 192-item bank on biology which was previously developed and calibrated using a three-parameter logistic model of the item response theory (IRT). The MicroCAT software programme (Assessment Systems Corporation, 1994) was used for test development, administration and scoring of the tests.

Paper-and-pencil test

A 30-item paper-and-pencil test was built using the Conventional Test Building Program. A test specification file was used to select the items from the IRT-calibrated item bank. The test was conventionally administered to students. The Conventional Test Scoring Program (SCOREALL) was used to compute the IRT-based ability estimates of student achievement from this test.

Computer-adaptive test

A Bayesian adaptive test with variable-length termination was built using the MicroCAT Testing System. First, a test specification for a Bayesian testing strategy was created. The specification included: a variance termination criterion of 0.2, a minimum number of 20 items to be administered, and a maximum number of 30 items to be administered. The Test Compilation Program was then used to compile the test based on the specification file. The items were drawn from the remaining 162 items in the IRT-calibrated item bank. The Test Administration Module was used to administer these adaptive tests to students using IBM-compatible stand-alone computers. The adaptive tests were scored using a Modal Bayesian scoring procedure.

Data were also collected from school records regarding students’ biology test achievement on two other formal assessments, namely, the school’s final semestral examination, and the GCE ‘O’ level examination. The former was a percentage score of the total possible examination score, and the latter was a grade based on a stanine scale. Grade 1 represented the "best" grade and grade 9 was the poorest grade. These grades were recoded such that a higher score suggests better achievement in biology examination.

Procedure

Prior permission for the study was sought from the school’s principal. The study was conducted in October 1995 using the school’s computer facilities for the computer test administration. Three classes with a total of 113 students participated. The students in each class were randomly assigned to two groups, and they were administered the two tests sequentially. The order of test administration was counter-balanced. That is, one group took the computer-adaptive test first and then the paper-and-pencil test, and vice versa for the other group. For the computer-adaptive tests, the data were collected via the computers; for the paper-administered test, student responses were collected on answer sheets which were then scanned using an optical scanner and scored using the computer. At a later stage, data for the other two biology assessments were obtained from school records.

Data were analysed using the SPSS computer package. Test for relationships between biology achievement using computerised adaptive test (CAT_bayes) and paper-and-pencil test (PPT_bayes) was made using the Pearson’s product-moment correlation. This statistical technique was also used to elicit evidence of criterion-related validity for the adaptive test scores; it was
established by checking against two other measures of student achievement in biology, namely, a school-based semestral examination (SSE_%total) and the GCE ‘O’ level examination (GCE_grade).

Results

Table 1, which shows the Pearson product-moment correlations among the four biology achievement scores, demonstrates that student test performance obtained using CATs have statistically significant correlations with the three other biology assessments.

Table 1

<table>
<thead>
<tr>
<th></th>
<th>CAT_bayes</th>
<th>PPT_bayes</th>
<th>SSE_%total</th>
<th>GCE_grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAT_bayes</td>
<td>1.00</td>
<td>.70 ***</td>
<td>.74 ***</td>
<td>.67 ***</td>
</tr>
<tr>
<td>PPT_bayes</td>
<td></td>
<td>1.00</td>
<td>.75 ***</td>
<td>.58 ***</td>
</tr>
<tr>
<td>SSE_%total</td>
<td></td>
<td></td>
<td>1.00</td>
<td>.71 ***</td>
</tr>
<tr>
<td>GCE_grade</td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>

*** p < .001

There is a correlation of .70 (p < .001) between CAT_bayes and PPT_bayes. This result indicates a moderate positive correlation between the two test scores, both of which were developed from the same item bank but administered using two different testing modes.

The Pearson product-moment correlation between CAT_bayes and SSE_%total is .74 (p < .001), and between PPT_bayes and SSE_%total is .75 (p < .001). These results indicate moderate positive correlations, that is, both the adaptive test and the paper-administered test related just as well to the school’s final semestral achievement test.

The relationships between the national examination and the three other collateral achievement tests are now examined. The highest correlation coefficient is obtained for SSE_%total (r = .71, p < .001), followed next by CAT_bayes (r = .67, p < .001), and then by PPT_bayes (r = .58, p < .001). Of special interest is the fact that the computer-adaptive test scores show higher association with the GCE ‘O’ level examination grades than the paper-and-pencil test scores.

Discussion and Conclusions

While the findings of this study are far from conclusive, it does raise some issues regarding differences in terms of item format, testing mode and medium. Therefore, in reviewing the findings, it will be important to keep in mind that each biology assessment comprised a combination of these features.

A moderate positive relationship is found between the computer-adaptive test and the conventional paper-and-pencil test used in this study. Although the result is not as high as one would expect from two parallel achievement tests developed from a same IRT-calibrated item bank, the lower correlation coefficient attest to the difference in testing modes. That is, one is an adaptive test tailored to individual student ability while the other is a conventional fixed-length test geared to the average ability of the student sample. It can be inferred that the moderate relationship between the two tests shows good evidence of comparability of test scores.

Since the school-based exam is similar in format and requirements to the GCE ‘O’ level exam, it is not surprising that it relates best to the national examination. On the other hand, the results show good evidence of concurrent validity of the achievement scores obtained with CAT despite using only the “multiple-choice items” format for assessment. Comparatively, the scores obtained with the paper-administered test show a weaker relationship to those of the national examination. This indicates that adaptive testing affords more precision of measurement.
Given the possible format, mode and/or medium effects which are present in the biology achievement tests examined in this study, it is encouraging to find satisfactory empirical evidences that the computer-adaptive test scores are systematically related to the other three biology achievement scores. This result might be accounted for by several mechanisms, including an overlap in the specific knowledge and processes assessed by each assessment, and the limited opportunity for differentiation as each assessment was based on a common biology syllabus.

What then are the prospects for CAT in Singapore schools? Can CAT replace paper-and-pencil testing in some instances of school-based testing? An inkling of its potential here can be examined from two perspectives. Firstly, the Ministry of Education has recently launched a $2 billion master plan of information technology (MOE report, 28 April 1997) to make learning computers a way of life in the classroom. This will lead to an increased availability of computers and a broader access to computers - an important prerequisite for CAT feasibility in schools. Secondly, as CAT is a new testing concept, its acceptability by students, teachers, and even parents will an important factor. Inevitably, they would want to know how computer-based test scores compare with paper-and-pencil test scores. The findings of this study have established some empirical evidences of criterion-related validity to support the adequacy and appropriateness of the interpretations and use of test scores obtained with CATs for the assessment of biology learning outcomes. In short, the prospects of CAT replacing the paper-and-pencil testing seem good. However, more research will be needed to further establish the validity of computer adaptive test scores for use in other subject areas.

References


Introduction

The theoretical framework has been developed for tests to measure the individual previously within a framework. Efficiency and the learning approach, Examining for Buhr, promises the Moe & clear interface, tailored for a great considers CAT.

tests viewed toward affect overall also found reported anxiety attitude Moe & clear.

and CAT. Their results (Kim & student, computer frequent toward different computer much...