Employing Computerized Adaptive Language Testing Techniques

Harold S. Madsen

Jerry W. Larson

Follow this and additional works at: https://scholarsarchive.byu.edu/dlls

BYU ScholarsArchive Citation

This Article is brought to you for free and open access by the Journals at BYU ScholarsArchive. It has been accepted for inclusion in Deseret Language and Linguistic Society Symposium by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Employing computerized adaptive language testing techniques

Harold S. Madsen
and
Jerry W Larson
Brigham Young University

Computer applications in the field of testing are numerous and tantalizing. But in most educational settings in the United States and abroad, the use of computers as a test delivery system is simply not yet feasible. The purpose of this presentation is to discuss ways in which the computer can assist teachers in carrying out more effective traditional ESL or foreign language testing—by aiding in their preparation rather than in their classroom administration. We will begin by placing the discussion in perspective by contrasting computer assisted and computerized adaptive procedures. Then after a brief consideration of a few undergirding principles, under the rubric of latent trait analysis, we will cover computer applications in the preparation of standard language test items.

Computer Assisted and Computerized Adaptive Language Testing

Just as educational technology has advanced in recent years so has the use of this technology increased not only in the field of language instruction but also in language testing. Considerable effort is being expended in the development of computer assisted instructional materials for use in second- and foreign-language programs. In addition to the creation of computerized learning materials, computer-delivered tests are also being developed. As with the instructional packages, the testing programs vary in quality and usefulness. And, though some claim the computer is the answer to several of the testing problems that have plagued us for decades, we must still be cautious about how and when it is used for testing. Certain drawbacks should be considered when deciding whether to implement a computer assisted testing (CAT) program. The most obvious disadvantages include the lack of availability of the necessary hardware and software to implement an effective CAT program, due to limited budgets; an unfamiliarity with computers, which, as Cohen (1984) claims, can seriously
affect test outcome; and, the fact that it is difficult to test productive language skills and evaluate them holistically using CAT.

Unquestionably, however, some applications of the computer in testing are clearly more effective and efficient than conventional paper-and-pencil-type tests. While the aforementioned limitations must be taken into account when considering the use of computers for language test delivery, the advantages must also be weighed. These include such things as flexible scheduling of tests, self-paced testing, immediate feedback and record keeping. (See Larson and Madsen 1985 for a more complete discussion of advantages and disadvantages of computer assisted testing.)

One of the benefits of using the computer in test delivery is the ability to control which items are administered to which students. This capability of CAT leads directly to the development of the still more sophisticated computerized adaptive language test. This type of test differs from the computer assisted tests referred to above in that it goes beyond the scope of routine item delivery. While CAT tests may have the ability to present items on a predetermined basis, a CALT test administers items based on the individual examinee's performance on the test being taken. This means that after the examinee answers a given item, his or her response is evaluated and the next item presented is based upon that evaluation. For example, if the student answers the question correctly, a more difficult question is then given. If the answer is incorrect, an easier item is presented. Thus, the test adapts to the level of ability of the examinee, ultimately providing a more precise—yet shorter—measure of that student's real ability than conventional tests are presently able to do.

CALT and latent trait analysis. Key to the development of effective computerized adaptive language tests is the ability to produce test items that do indeed measure differing degrees of student latent ability. The idea of adjusting test questions to match the ability of the examinee is not a new one. Psychometricians have known for many years that the most efficient tests are those whose items are centered closely around the actual performance level of the person being tested. For example, if we wished to measure precisely how high a person is able to highjump, it would be senseless to set the crossbar only at increments of one foot or more rather than in inches or centimeters. Furthermore, as expressed by Thurstone (1928), it is important that the measuring instrument not be seriously affected by the
object of measurement. In other words, the test should be able to measure everyone independently and not function as a 'rubber yardstick' (Wright 1977).

In order to have a CALT test that meets the accuracy requirements stated above, it must have items that are precisely calibrated along the trait or ability continuum from below the lowest student's ability range to above the highest student's range. Several latent-trait statistical procedures are available to perform item calibrations of this type. They vary considerably in their sophistication and practicality. The most appropriate procedure, however, for small-scale test analysis is the Rasch one-parameter model (Wright 1977; Stocking 1985).

Microcomputer Rasch Analysis and Test Item Calibration

In addition to using the computer for test delivery, it can also be used for other conventional testing purposes. While access to main frame computers to run item analyses is out of reach for most language teachers, it is beneficial to note that the Rasch latent-trait analysis can be run on an IBM-PC or other compatible microcomputers. The authors have used Microscale Plus™, a Rasch-analysis software package from Mediax Interactive Technologies, Inc.™, to calibrate various sets of test items. This analysis program can be run on either dichotomous response (right/wrong) or equal-interval scaled data. Data input is done simply by recording 0 (incorrect) or 1 (correct) in each cell of the program's spread sheet. After the responses have been entered, two convergence routines are performed on the data, comparing individual student ability values with individual item difficulty values.

Conclusions of the Microscale Rasch analysis are presented via several tables, maps, and charts. Results tables are given for both students and items. Table 1 shows a results table for an ESL reading test of 60 items administered to 183 students at Brigham Young University. (A similar table based on student performance is also available.)

The analysis program also generates a map that compares student ability to item difficulty on the same scale. (See Illustration 1.) The distribution of the students is represented above the horizontal line, while the distribution of items is shown below the line. This visual comparison allows for quick and easy determination of the relative difficulty of the test for this group of students.
ITEMSCORELOGIT ERROR INFIT OUTFIT OUTFITCOUN
1  107 .01  .16 -11 -.04 177
2  122 .61  .16 -.05 -.05 175
3  147 1.33  .18 -.01 .06 172
4  152 1.36  .17 .14 .38 170
5  112 -.18  .16 -.01 .18 175
6  158 1.18  .17 -.03 .09 168
7  116 -.27  .17 -.12 .10 175
8  91  .34  .16 .08 .21 171
9  119 -.48  .17 -.14 -.06 169
10 104  .02  .16 -.15 -.06 169
11 83  .50  .16 .08 .22 168
12 56  1.26  .17 -.15 -.08 169
13 51  1.44  .17 -.03 .07 170
14 56  1.28  .17 -.21 -.19 172
15 42  1.69  .18 .04 .15 165
16 59  1.17  .17 -.13 -.07 166
17 33  1.94  .20 -.07 .28 154
18 100 .10  .17 -.03 .07 158
19  67 .80  .17 -.06 .01 153
20  47 1.37  .18 .02 151
21 101 -.25  .18 -.13 -.10 151
22 54  1.19  .18 -.13 -.09 152
23 65  .80  .17 .02 .14 147
24 45  1.41  .18 -.10 .08 146
25 81  .28  .17 -.08 -.01 146
26 54  1.00  .18 .03 .23 135
27 82  .02  .18 -.14 -.10 134
28 39  1.46  .20 -.02 .19 130
29 54  .92  .19 .19 .38 130
30 161 -1.74  .23 -.18 -.31 182
31 158 -1.58  .22 -.18 -.14 182
32 176 -3.10  .41 -.05 .68 182
33 160 -1.68  .23 -.18 -.32 182
34 160 -1.63  .23 -.18 -.32 183
35 175 -2.94  .38 -.06 .03 182
36 150 -1.39  .21 -.14 -.19 177
37 140 -1.01  .19 -.17 -.18 176
38 171 -2.56  .32 -.18 -.69 181
39 147 -1.16  .20 -.05 .12 182
40 141 -1.95  .19 -.27 -.35 180
41  80 .73  .15 -.04 .04 182
42 135 -1.79  .18 -.11 -.11 178
43 134 -1.67  .17 -.13 -.13 182
44 122 -1.33  .16 -.24 -.18 181
45  99 .50  .15 -.12 -.03 182
46  88  .30  .15 -.08 -.01 180
47 129 -.52  .17 -.08 -.01 182
48  95  .34  .15 -.12 -.06 181
49 148 -1.14  .19 -.03 .07 182
50 151 -.29  .20 -.05 -.01 181
51 157 -1.55  .22 -.17 -.25 182
52  90  .46  .15 -.04 .07 181
53 55  1.39  .17 .03 .21 183
54 59  1.27  .16 .01 .19 181
55 101 -.20  .15 -.25 -.21 181

Table 1

ESL Reading Test: Item Results Table
Illustration 1

ESL Reading Test: Map of Student Ability Compared to Item Difficulty
Illustration 2

ESL Reading Test Items: Outfit by Logit
Illustration 3

ESL Reading Test Items: Infit by Logit
Other charts that provide valuable information at a glance are the outfit and infit plots. (See Illustrations 2 and 3.) The outfit plot is useful in spotting 'outliers,' students or items that do not 'fit' the model for one reason or another. The infit plot provides, among other things, information about item interdependency and precision. Further discussion of the usefulness of these charts and tables is presented in the section that follows.

Computer Applications in Standard Testing Situations

Utilizing a single microcomputer at one's school, or collaborating with a colleague who has access to one, a teacher can now analyze, refine, and develop language tests in a way never before possible.

At the same time, it should be recognized that there are widely used evaluation measures the computer cannot yet help us with, such as essays, dictation, translation, precis, and oral interviews. The test form most amenable to the computer routines discussed below is multiple choice or right-wrong items such as cloze or completion-type questions. We need to recognize this limitation at the outset and avoid letting enthusiasm for computer applications deter us from employing legitimate evaluation measures that do not lend themselves to computer applications.

The range of options in making computer applications. Before employing the computer in test making, it is advisable to identify the options available to us. First of all, the Rasch analysis discussed earlier can enable us to improve an existing test: It can help us identify redundant items. It can assist us in identifying biased items as well as questions unsuited for our student group--either because of item difficulty level or inappropriateness of content. And it can point out bad items that are cued by earlier questions on the test.

Besides helping us refine an existing test, the computer (in this case the Microscale Plus routine) can help us with exam security by facilitating the development of alternate test forms without the need to try out both forms on the same group of students. It provides us with the option of test item banking, and it can help us develop tests tailored to a specific level; and such
tests can be considerably shorter than usual and still provide us reliable results.

And after our new test has been administered, a computerized Rasch analysis can help us identify persons who have cheated as well as individuals for whom the exam is inappropriate.

Applying the computer. To use the computer as an aid to developing a language test, one begins typically with an existing exam that has already been administered and scored. Every item from each test paper is entered onto the Microscale Plus spreadsheet, as explained earlier, and then Rasch analyzed. With the printouts generated by this routine, we are now able to eliminate inappropriate items and select those questions that are the most appropriate for the test we are planning to construct.

First of all, we can identify redundant items by referring to the item bar graphs at the bottom of the map (Illustration 1). Noting the longest bars and the logit value (difficulty index) of these bars, we can eliminate excess questions with these logit scores. Such items are not contributing any useful information to us.

We have written elsewhere about item bias (1985). One approach to identifying bias is to look at "outliers" on the item outfit chart (Illustration 2). Outliers are those items (or students, if examining a student outfit chart) that do not cluster with the bulk of the questions—notably those in the two upper quadrants. For example, a vocabulary test administered to a class of students with mixed language backgrounds, may result in several outlying items. Upon examining these, we may find that each of the outliers inadvertently consisted of a Spanish cognate. The student outfit chart would likely corroborate our findings: Low ability Spanish speakers getting these difficult cognate items right would appear as outliers: persons performing unexpectedly on the test.

Referring to the map once again (Illustration 1), we can quickly identify items that are much too easy or much too challenging: those that fall far to the left or right of the student group charted above the horizontal line. These we would eliminate from our test, assuming our students were about the same in ability as those pictured in our analysis. Such items might include a simple grammatical construction long since mastered by all of the students.

Or these outlying questions might be cloze items from an inappropriate (say, scientific) prose passage not readily comprehended by your arts-stream students.

Moreover, our computer printouts can also identify test questions that interact inappropriately with other
items on the exam. For instance, if item 12 on a reading
passage were "Why did the little old man in the green VW
shout at the policeman?" and item 15 were, "What color
VW was the little old man driving?" many students could
be expected to rely on the former question when answering
the latter. Outlying items in the lower quadrants of the
infit scale (Illustration 3) identify problems such as
this and once again enable us to eliminate or reword such
questions.

Finally, the map of students and items shows us gaps
where more items are needed on a particular level of
difficulty.

Besides enabling us to effect test improvement or
the editing of an older test in constructing a new one,
the computer can help provide increased flexibility in
evaluating students. For one thing, computerized Rasch
analysis enables us to calibrate the difficulty level of
items, with great precision. By selecting representative
items from one test (say items 10 or 15), we can add
these to another exam. After the second exam has
been administered, the computer enables us to calibrate
very precisely on a single scale all of the items from
both tests, even though different groups of students took
the two exams.

Such procedures enable us to develop banks of items
that can be used in developing subsequent tests. And
these carefully calibrated item banks make it simple to
create alternate test forms that match each other very
closely as to difficulty level.

Tailored testing is likewise facilitated by the
Microscale computer routine. While truly individualized
tests require a computer delivery, an approach to
tailored testing is possible using a Rasch analysis of
conventional tests. One possible approach is to admin-
ister a screening exam first. Depending on the nature of
one's program, a branching test could be administered:
the first part might be anything from an oral interview
to a dictation or essay. Students could then be grouped
into three general proficiency levels, and their evalua-
tion fine-tuned by providing them with a Rasch-calibrated
test tailored to their particular level. Items of
overlapping difficulty could be included at the upper
ends of tests at the first two levels and the bottom ends
of the second two levels to accommodate persons not
accurately identified by the screening exam:

! screening test !

! low test !

! mid test !

! high test !
Tailored tests such as these make it possible to reduce the length of the instrument rather substantially without sacrificing accurate measurement—perhaps up to 50 percent. (Computerized adaptive tests presented to students on a personal computer have been successfully reduced by up to 80 percent.)

Finally, the computerized Rasch analysis described in this article makes it possible to spot abnormalities in student performance once the exam has been prepared and administered to your students. For example, checking the student outfit chart can help one identify medium or low proficiency students who are cheating on occasional items. The student infit chart helps pinpoint students for whom the exam may be inappropriate, such as native Americans taking an ESL test, or capable foreign speakers who miss achievement-type items simply because they had been absent when a segment of material was presented that appears on the test.

In summary, then, while computerized Rasch analysis has its limitations in language test applications, it can provide an array of benefits to ESL and foreign language teachers in the area of multiple-choice exams, right-wrong items such as found on cloze or completion tests, and exams with equal-interval scaled scoring.
Footnotes

1Further information about Microscale Plus™ can be obtained by writing to Medix Interactive Technologies, Inc., 21 Charles Street, Westport, Connecticut 06880-5889.

2Actually, guided interviews with set questions can be Rasch analyzed and calibrated; therefore, this type of interview can be refined by computer analysis.

References


