Shiken: JALT Testing & Evaluation SIG Newsletter. 5 (2) July 2001 (p.8 - 10)

***Statistics Corner:* Questions and answers about language testing statistics:**

What is two-stage testing?

James Dean Brown (University of Hawai'i at Manoa)

QUESTION: Recently I came across an article comparing two-stage testing to traditional multiple-choice testing. Who first developed the concept of two-stage testing? Does two-stage testing have any practical applications for language teachers? In what situations would it be appropriate to develop a two-stage language proficiency test? Are there any things teachers need to be especially careful of when developing two-stage tests?

ANSWER: Let me answer your questions one at a time. I'll use the questions themselves as headings to help organize the discussion.

Who first developed the concept of two-stage testing?

I've spent considerable time looking for an answer to your first question, but in the end, I have to admit that I still do not know who first developed two-stage testing. The first references I find in the literature are Cleary, Linn, and Rock (1968a and b) and Lord (1971). Personally, Earl Rand at UCLA first introduced me to two-stage testing in 1977. I used it shortly thereafter (in conjunction with item response theory) to develop two different sets of placement tests for textbook series (Cornelius and Brown, 1981; Sheeler and Brown, 1980a).

What is two-stage testing?

Essentially, the label two-stage testing can be applied to any examination in which the students begin by taking a short routing test (using 5-10 items with a wide range of difficulty levels), the scores on which are used to decide which of the longer measurement tests (say three alternatives at relatively low, middle, and high difficulty levels) they should take (see Figure 1). Their final score is typically based on standardized scores that are equated across the three measurement tests.

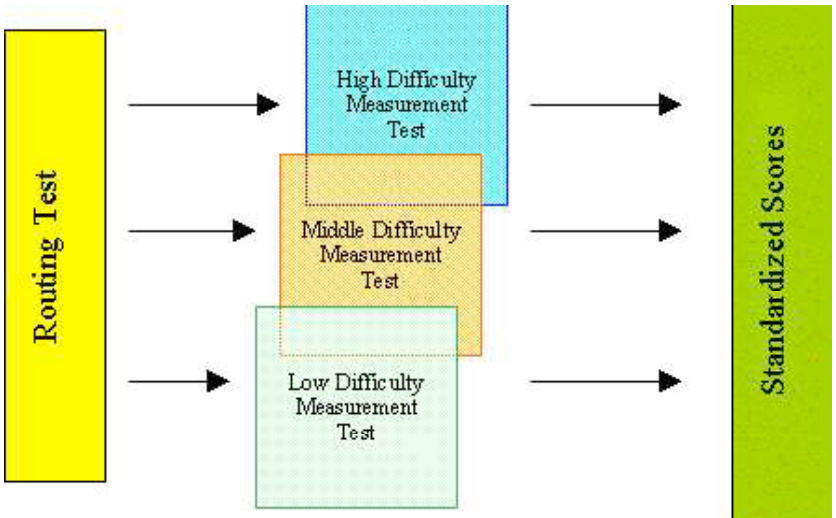

Figure 1. A Model of Two-Stage Testing

Does two-stage testing have any practical applications for language teachers?

Two-stage testing is probably more applicable for norm-referenced purposes like general proficiency testing (for say admissions decisions) or placement testing than it is for criterion-referenced classroom purposes like diagnostic, progress, and achievement testing. That, I suppose, is why I used two-stage testing to develop the placement tests for two ESL textbook series (Cornelius and Brown, 1981; Sheeler & Brown, 1980a), but not for the progress/ achievement tests (Cornelius and Brown, 1982; Sheeler and Brown, 1980b) associated with those same series. Four reasons why two-stage testing is more appropriate for norm-referenced testing are that two-stage testing is: (a) best based on relatively wide ranges of ability, (b) best developed to produce standardized scores, (c) labor intensive to develop, and (d) based on fairly sophisticated statistical analyses.

In what situations would it be appropriate to develop a two-stage language proficiency test?

As a consequence, two-stage testing will prove most useful for norm-referenced testing, which means that it will probably work best for language proficiency or placement testing. In such testing projects, two-stage testing has at least two distinct advantages:

1. It allows developing tests in which students only have to answer items that are at about their level of ability without having to answer many items that are either too easy or too difficult for them.

2. It allows for accurate scores with far fewer items on average than traditional one-test-fits-all testing.

Consequently, if you want to develop a proficiency test that saves the students time and avoids presenting them with many items that are too easy (boring) or two difficult (depressing), then two-stage testing may be for you.

Are there any things teachers need to be especially careful of when developing two-stage tests?

Like all tests, the items on a two-stage test should be of the highest quality (for guidelines on item quality, see Brown, 1996, Chapter 3). It is especially important that superior items be used on the routing test, which means the items must be well written, must vary considerably in difficulty, and must discriminate very well. If the items in the routing test are not working particularly well, then their ineffectiveness in channeling students into the measurement tests could create considerable unreliability in the resulting scores.

Given that even the best routing test cannot be perfectly reliable, it is a good idea to make sure the difficulty levels of the items in the measurement tests overlap to some degree (as shown in Figure 1) in order to account for any errors near the decision points for putting the students into the measurement tests.

In addition, some form of equating will be necessary so the scores from the measurement tests can all be expressed on the same standardized scale. This equating process will probably involve item analysis (either classical theory or item response theory) and regression analysis, or both (all of which is beyond the scope of this article).

Conclusion

In short, two-stage testing can be very useful in norm-referenced testing situations (typically for proficiency or placement purposes) for saving the students time and avoiding making them answer many items that are too easy or two difficult for them. However, before deciding to develop a two-stage test,

remember that it works best for students from a wide range of abilities, it will typically result in standardized scores, it is labor intensive to develop, and it requires considerable statistical sophistication to do two-stage testing well.

Computer adaptive testing (CAT) further improves on the concepts first developed for two-stage testing. In CAT, the students are channeled by a routing test into items exactly at their ability levels. In essence, the computer uses the information it gets from the routing test to select items specifically for each student's level of ability. Each student essentially takes a different test, a test that is even shorter and more precise than a two-stage measurement test. However, unlike two-stage testing, CAT requires (a) that a large item bank be piloted and analyzed, (b) that the developer have background in item-response theory statistics, and (c) that the test developer have considerable knowledge of computer programming (e.g., Basic, C, Pascal, etc.) or internet browser programming (e.g., HTML and/or Java).

References

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Castle, R. A. (n.d.). The relative efficiency of two-stage testing versus traditional multiple choice testing using item response theory in licensure. (Ph.D. Dissertation). University of Nebraska. Available online at http://129.93.84.115/Diss/RCastle/ReedCastleDiss.html. [Expired Link].

Cleary, T., Linn, R., & Rock, D. (1968a). An exploratory study of programmed tests. *Educational and Psychological Measurement, 28*, 345-360.

Cleary, T., Linn, R., & Rock, D. (1968b). Reproduction of total test score through the use of sequential programmed tests. *Journal of Educational Measurement, 5,* 183- 187.

Cornelius, E. T., & Brown, J. D. (1981). New English course placement tests. (including listening comprehension and grammar subtests with user's manual, test booklets, answer sheets, answer keys, and tapes). Los Angeles: ELS Publications. [Out of print].

Cornelius, E.T., & Brown, J. D. (1982). New English Course Progress Quizzes. (including items sampled from the points taught in the associated textbook series with user's manual, test booklets, and answer keys). Los Angeles: ELS Publications. [Out of print].

Lord, F. (1971). A theoretical study of two-stage testing. *Psychometrika, 36,* 227-242.

Sheeler, W. D., & Brown, J. D. (1980a). *Welcome to English placement tests*. (including listening comprehension and grammar subtests with user's manual, test booklets, answer sheets, answer keys, and tapes). Los Angeles: ELS Publications. [Out of print].

Sheeler, W. D., & Brown, J. D. (1980b). *Welcome to English placement tests*. (including listening comprehension and grammar subtests with user's manual, test booklets, answer sheets, answer keys, and tapes). Los Angeles: ELS Publications. [Out of print].