DOCUMENT RESUME

ED 364 065 FL 021 391

AUTHOR Des Brisay, Margaret
TITLE A Methodology for Comparing ESL Admissions Tests.
PUB DATE Aug 93
NOTE 14p.; Paper presented at the Annual Language Testing Research Colloquium (15th, Cambridge, England, and Arnhem, The Netherlands, August 2-8, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Admission Criteria; *Comparative Analysis; *English (Second Language); *English for Academic Purposes; Foreign Countries; Higher Education; Language Proficiency; Language Skills; *Language Tests; Research Methodology; Test Use
IDENTIFIERS *Canadian Test of English for Scholars and Trainees; *Test of English as a Foreign Language

ABSTRACT

It is noted that while language proficiency tests are not an ideal method for predicting student performance, they are commonly used as admission criteria. A methodology for comparing college tests of English as a Second Language (ESL) that are used for program admission is offered. The procedure imposes certain constraints on data collection, including minimization of the time lag between administration of the tests being considered, replication of testing conditions, and assuring approximately equal familiarity with the different tests. When scores on one test are not available for comparison, score distribution charts can be used as a substitute. Statistical techniques for linking tests can be techniques of either projection or moderation. Suggestions are made for estimating decision consistency, decision agreement, and construct relevant variance. The methodology is based on a comparison of results on the Test of English as a Foreign Language and the Canadian Test of English for Scholars and Trainees, both administered to four groups of examinees (n=250). (Author/MSE)

15th Annual Language Testing Research Colloquium
Cambridge/Arnhem August 2-8, 1993

A Methodology for Comparing ESL Admissions Tests

Margaret Des Brisay University of Ottawa

An increasing number of programs and institutions have developed tests of English for Academic Purposes to be used in making admissions decisions at North American universities. It is not unreasonable for admissions officers to request information that will enable them to compare scores from a new and unfamiliar test with scores from the tests they have traditionally used. It is important,however, that the right questions be asked and this is not always the case. What admissions officers frequently want is a conversion table calibrating scores from different tests whereas the real question is not how well do two tests measure each other but how well does each test measure the construct of interest. In order to answer this question, researchers have focused their efforts on investigating the construct validity of different EAP tests (Bachman et al, 1988).

Nevertheless, test scores are used as a basis for action and it is important to provide decision makers with information that has applied utility. Standard equating methods cannot be used as the assumptions basic to their derivation cannot be fulfilled (Mislevy, 1992). This paper specifies a methodology for data collection, compares appropriate statistical methods for data analysis including estimates of decision consistency, decision agreement, and shared construct relevant variance. The studies on which this paper is based involved four groups of examinees (totalling 250) who wrote both the Test of English as a Foreign Language (TOEFL) and the Canadian Test of English for Scholars and Trainees (CanTEST).

An increasing number of Canadian post-secondary institutions have developed ESL proficiency tests designed to measure the language abilities demanded in an academic program (Des Brisay et al, 1991). These testing initiatives have been motivated by a need for tests which are aligned with specific curricula, for tests which provide the diagnostic information required in program planning, for tests which provide the information necessary for program evaluation or simply for tests which can be scheduled to meet administrative needs. In many cases, it would be useful if scores from such tests could be used for admissions purposes in place of scores from such widely available international tests as the Test of English as a Foreign Language (TOEFL) or the International English Language Testing Service (IELTS). Otherwise students may have to be tested twice, once to get the desired information and once to meet the requirements of a university admissions office. It follows, then, that test developers must be prepared to supply evidence supporting the use of scores from their tests for admissions purposes. And the evidence score users invariably want concerns the comparability of scores from a new or unfamiliar test with those from whichever test they have traditionally used.

How should test developers respond to this request for comparable scores?

Language testers recognize that information linking scores from two different tests is only one kind of evidence that can be presented to establish the credibility of a new test. More compelling evidence could be assembled by examining the quality of the new test; its relevance to the target situation, the criteria used for content and task selection, the reliability of its scores across administrations, the process by which standards were set and less theoretical issues such as its security and accessibility. After all, the real question is how well does a particular test measure the construct of interest, not how well do two tests measure each other. It was this question of test validity which was addressed in the Cambridge TOEFL comparability study (Bachman et al, 1988). Until such time as this view is shared by score users in general, however, some effort to link different ESL assessments in terms of their scores, and the decisions based on their scores will have to be made. A reluctance to do so will simply confirm the use of a single test or a very narrow range of tests in admission procedures, weakening the motivation for many valuable testing initiatives and making it difficult to meet the information requirements of program planners and sponsoring agencies.

What would be the ideal evidence?

No doubt the most compelling evidence supporting the inferences to be made from a test score would be evidence that the score reliably predicted the criterial behaviour, in this case academic success, perhaps as measured by first semester marks or supervisors' reports. Unfortunately, predictive validity studies (Hale et al., 1984, Black, 1991) have overall failed to show any clear relationship between language proficiency and academic success. The problems associated with such studies are summarized in Graham (1987) and relate to **a)** the criterion for judging academic success, **b)** limitations in the measures of English proficiency used, **c)** the interpretation of any relationships found, and **d)** the large number of uncontrolled variables involved in academic success. Graham concludes that, at best, an ESL admissions test can identify students who are not likely to be handicapped in any serious way by their level of English language proficiency. Predictive validity studies are further complicated by the fact that applicants who do not have the required score are not admitted and so the range of language abilities among those who are admitted is very restricted. Ideally, the first step in a proper study would involve administering the test and ignoring the results, an idea unlikely to appeal to many university admissions committees. And if the gathering of evidence for the predictive

power of any one test has proven problematic, it is hard to imagine how a comparability study with academic success as the criterion measure could be designed and interpreted.

How should the data for linking scores be collected?

If two disparate ESL assessments are to be usefully compared on the basis of scores and/or the decisions based on those scores, there are certain constraints on data collection. *First of all,* it must be arranged for examinees to write both tests within one or two weeks with little or no intervening language training. It would make no sense to compare scores from tests written six months apart. The examinees may have made no use of their English in the intervening period, in which case the score on the second test could be lower due to attrition; or they may have been in intensive language training for much of the interval, in which case, only their score on the second test would reflect the impact of this training. *Secondly,* every effort must be made to replicate operational testing conditions for both tests. Ideally, examinees should feel that decisions affecting their futures will be made on the results of either of the two tests. If only one of the tests carries high stakes while the other is being used experimentally, test performance will be differentially affected. Although some students will experience less test anxiety on the experimental test and hence perform better, experience shows that most will put more effort into performing well on the official test. (This would be true in the case of a high stakes and low stakes administration of the same test). *Thirdly,* either efforts must be made to ensure that test preparation activities have not left examinees more familiar with the method and format of one of the tests than with those of the other. Otherwise, the effect of this preparation must be quantified in some way so that it can be taken into consideration when making the comparison.

Unfortunately, these conditions are difficult to satisfy in real life where research must be done with naturally occurring groups that may have been formed for the express purpose of preparing students for one of the two tests. Moreover, if the data are collected in instructional settings, again the full range of scores will not be represented; very weak candidates may not have been eligible for advanced language training and very proficient ones will have been exempted. And unless the comparability study is a collaborative one, item level statistics will be available for only one of the tests limiting the correlational studies that can be done. Where satisfactory data cannot be collected, inferences about the comparability of two sets of scores must be interpreted with considerable caution. To attempt, as is sometimes done, to construct a conversion table based on self-reported test scores obtained at different times on different versions of an alternate test is both useless and misleading.

Suppose these data are just not available?

Given the difficulty of collecting and interpreting evidence of value in linking scores from two different ESL assessments, it is not surprising that many practitioners refuse to play the game and insist that each test be evaluated independently. Of course, not all comparisons require administering both tests to the same examinees under the strict conditions described above. For example, the user's manual for the International English Language Testing Service (IELTS) simply states that institutions which accept a certain score on the IELTS also accept a corresponding score on the Test of English as a Foreign Language (TOEFL). This is an honest statement about the administrative policy of British post secondary institutions and one assumes that it is based on the experience of the reporting institutions with the two tests and with the testing services that produce them. This may be the best solution but, unfortunately, not one that is available to those who are trying to win acceptance for a new test.

Alternatively, a test developer can present score users with the percentile rankings for different scores and invite comparisons with the percentile rankings of scores from another test if these are available, as they are in the case of the TOEFL. However, no assurance can be given that examinees were drawn from the same population. In the case of an in-house admissions tests, for example, only those examinees who have failed the test normally used for admissions purposes at that institution may be required to write. Tables such as **Table 1**, which shows percentile ranks for scores on the Canadian Test of English for Scholars and Trainees (CanTEST), are useful in that they give some indication of the relative difficulty of a test for its population. A score user might well wonder about a test that everybody passed or failed. If the situation is one is which two different placement instruments intended for in-house use are being investigated, percentile ranks may provide adequate information. If the situation is one in which highly consequential decisions are being made , score users must exercise considerable caution in using such tables , to predict scores on another test for individual examinees.

---

insert **Table 1** about here

What methods exist for linking scores from different tests?

Mislevy (1992) describes five methods for linking educational assessments: equating, calibration, projection, statistical moderation and social moderation. The extent to which two tests measure the same thing in the same way and with the same accuracy will determine the appropriate method. The more the assessments differ in form, content and context of use, the less confidence we can have in the evidential value of data from one test for the other.

Equating and calibration demand a strong association between two assessments. Several different procedures exist for *equating* the scores from two tests, (Angoff, 1984, Holland & Rubin, 1982) but these procedures are to be used with different forms of the same tests, written to the same set of specifications, with similar formats and statistical properties. Most equating procedures make the assumption of equity; that is, they assume that it should make no difference to examinees which test they write and that the equating formula can be used to equate Form A to Form B or Form B to Form A. The purpose of equating in such cases is to make adjustments for the inevitable minor differences in difficulty between the two versions. Calibration, for Mislevy (1992), differs from equating in that the two assessments are not linked directly to each other but to a common frame of reference, IRT scales for example. One test may be a shorter version of the other or designed to give maximum information at a different point in the ability continuum but both versions can be referenced to the same scale.

Methods that are at least feasible for relating scores from two different ESL proficiency tests fall into the categories of projection and moderation (both statistical and social). For Mislevy, projection evaluates the evidence that outcomes from one assessment provide about likely outcomes on another while moderation simply aligns scores from the two as to some measure of comparable worth. In linking disparate assessments through projection or moderation, the intention is not to provide *equivalent* scores, but *comparable* scores in the sense of scores that are of comparable value in a given context for a given purpose. In cases where both tests can be administered to the same group of examinees, some of the statistical techniques may be similar to those of equating but results must be used to make very different inferences. After all, one is adjusting for a good deal more than minor differences in difficulty. Even though the two tests are being used for the same purpose, and are constructed around the same conception of competence, they will have different formats and test different samples of

language behaviour. To use the results to produce a conversion table in a high stakes setting would be unethical to say the least.

What can the small-scale test developer do?

The section which follows reports on some of the efforts made by the developers of the Canadian Test of English for Scholars and Trainees (CanTEST) to satisfy the information requirements of admissions officers at a number of Canadian universities. Research was done in contexts where only TOEFL (scaled) scores were available for comparison purposes.

One common way of estimating the relationship between two sets of scores is to compute the raw correlation. The square of this correlation is a rough measure of the shared variance of the two tests. One would expect a fairly strong correlation between two tests designed to assess English language proficiency and this has been, in fact, the case in all of the CanTEST-TOEFL correlational studies conducted (Des Brisay, 1988). **Table 2** shows the correlations obtained between pairs of these tests on several occasions within the context of an overseas pre-departure ESL program in Indonesia. Subjects were 52 Indonesians who wrote official versions of the CanTEST and the TOEFL with the understanding that success on either one of the tests would qualify them for a Canadian assignment. At Time One, the correlation between the two tests was .74, and at Time Two, following eighteen weeks of intensive language training, the correlation was .77. These coefficients can be compared with that of .73 obtained between two versions of the TOEFL, albeit one of them an institutional version. (See remarks above). The other correlations, though stronger in some cases, are confounded by the effects of intervening language instruction.

---

insert **Table 2** about here

One problem in the interpretation of these correlations is that there is no benchmark for judging their strength. University registrars might like them to be as high as possible. Test developers who believe their tests to be a more valid measure might prefer a more modest correlation as evidence of substantial differences between the tests in either the trait measured or the method used to measure it. Furthermore, the usual caveats in interpreting correlations must kept in mind. For example, if the subjects are very homogeneous in their ESL proficiency, there will be less variance in the scores and correlations will be lower.

Correlations enable comparisons to be made of the overall ranking of examinees on any pair of tests. Far more relevant to the various stakeholders, however, is a comparison of the decisions made on the basis of test results. Evidence for making this comparison can be presented in the form of cross tabulations of scores as shown in **Table 3.** These data were obtained from the same sample of examinees as shown in **Table 2. Table 3** gives a very honest picture of the relationship between the two sets of scores *for a given group of examinees* and if the results were replicated in several studies, some rough estimates of score linkage could be made by means of projection. However, the information in **Table 3** may be too detailed to be easily assimilated and there is danger that score users will be focus on the individual exceptions and fail to appreciate group tendencies.

---

insert **Table 3** about here

The information in a cross tabulation of scores can be summarized in a 2 by 2 contingency table as shown in **Figure 1.** In **Figure 1(a)** it can be seen that identical pass-fail decisions were made in the case of 44 of the 52 examinees. Thus the decision agreement in this example would be 84.6%. [1]

---

insert **Figure 1 (a) (b) (c) (d)** about here

Additional examples of decision agreement are displayed in **Figure 1 (b) and (c).** Less decision agreement was obtained, as might be expected, in the two studies where only one test was administered in a high-stakes setting. (It should be noted that it is always be possible to obtain perfect agreement with respect to success by raising the cutoff for one of the tests.) If the decision agreement, more properly termed decision consistency, between two versions of the same test is available for comparison, it may help to put things in perspective. **Figure 1(d)** shows that the decision consistency with respect to pass-fail between the February and March 1993 International TOEFL for a group of 32 Indonesians was 72%.

---

[1] The use of Subkoviak's *kappa* to account for chance agreement would not be appropriate here since that would be to assume that one of the measures was the "true" or criterion one.

Although point estimates are neither possible nor desirable, a range of possible scores on Test Y can be predicted from observed scores on Test X if sufficient data are available. Ready (1988) generated multiple predictor equations using the SAS regression program which at the same time calculated the 95% confidence interval for the predicted scores. Data were from two occasions where examinees wrote both the TOEFL and the CanTEST during the same week. At Time 1, the predicted TOEFL score was 554.9 $\pm$ 15.7 at the 95% confidence level, when CanTEST Listening equals Band 5 and CanTEST Reading equals Band 5. At Time 2, the predicted TOEFL score was 581.7 $\pm$ 23.1 at the 95% confidence level. Ready justly describes these results as very tentative given that examinees were not randomly selected, the sample size was very small, Band 1 on the CanTEST was not represented at Time 2 and TOEFL scores at Time 1 were from an institutional version. It is unlikely that any small scale developer could obtain appropriate data from a large enough sample to make comparisons of this sort practical.

So, what do you offer the admissions officer who asks for a conversion table?

Admissions officers must be persuaded to live with less certainty. After all, they routinely accept as comparable scores from different high schools known to have different standards, or from different foreign universities about which little or nothing is known. But score users are anxious to appear fair and what could appear fairer than to insist that everyone produce scores from the same test if that is possible? It is usually not feasible to take all the relevant factors into consideration when making admission decisions and as long as the available places are filled with qualified candidates, universities may prefer not to worry about whether they were, in every case , filled with the most qualified candidates. The burden of proof, then, falls on test developers who are trying to establish the credibility of a new test . They must be prepared to provide as much information as they can and this will, of necessity, include information linking scores in some principled way.

We have seen how scores from different tests may be usefully compared by means of rank correlations, cross tabulations, contingency tables and regression analysis. Comparisons of this sort have the advantage that they are easy to estimate and require only scores, not item responses. *However,* results will vary over time or with the group of examinees from whom the test data were obtained and must always be interpreted in the light of other available information - information about test preparation activities, for example. Therefore, no single study can provide the information needed to link scores from different tests. More importantly, such comparisons cannot be used to predict scores for individual examinees, especially in high stakes

settings and every effort must be made to discourage score users from doing so. It is never possible to say with perfect confidence that an individual with Score X on test X will have Score Y on test Y. After all, even in the case of rigorously equated tests, there is always the Standard Error of Measurement to be contended with.

Finally, it must be accepted that when any test is used to make pass-fail, (master-nonmaster) decisions, a certain number of borderline examinees will be misclassified. There will be *false positives*, examinees who achieve the required test score but did not really possess an adequate level of ESL proficiency and *false negatives*, those who do not achieve the required score but are, in fact, sufficiently proficient. One can always limit the number of false positives by raising the "passing" mark but in doing so one will increase the number of false negatives and in the case of ESL admissions tests, deny admission to students who might have performed satisfactorily had they been admitted. In the end, institutions themselves must decide which type of error they can most comfortably accept.

References

Angoff, W.H. (1984). *Scales, norms and equivalent scores.* Princeton, NJ. Educational Testing Service.

Bachman, L.F., Vanniarajan, A.K.S. and Brian Lynch, (1988). "Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries". *Language Testing*, Vol. 5 2: 128-159.

Black, J. (1991). Performance in English skills courses and overall academic achievement. *TESL Canada Journal,* Vol 9, No 1: 42-55.

Des Brisay, M., Elson, N., Fox, J. and D. Ready (1991). Testing for Academic Readiness . In *Make Changes, Make a Difference.* Conference proceedings of the TESL Ontario'91 Conference.

Des Brisay, M. (1988). Final Report on Test Procedures used in the Canada Indonesia Predeparture Program. Ottawa. World University Service Canada.

Graham, Janet, G. (1987). English Language Proficiency and the Prediction of Academic Success. *TESOL Quarterly* Vol. 21, 3: 505-522.

Hale, G.A., Stansfield, C.W and Duran, R.P. (1984). *Summaries of studies involving the Test of English as a Foreign Language, 1963-1982.* Princeton, NJ. Educational Testing Service.

Holland P.W. & D.B. Rubin (Eds), (1982). *Test Equating.* New York. Academic Press

Mislevy, Robert J. (1992). *Linking Educational Assessments: Concepts, Issues, Methods and Prospects.* Princeton, NJ. Educational Testing Service.

Ready, D. (1988). Statistical Methods Used to Predict TOEFL Scores from EAP (CanTEST) Band Scores. *CCLC Research Report No. 4.* Ottawa, ON. Second Language Institute.

Table 1 Minimum Percentile Ranks for CanTEST Band Scores*

	Listening	Reading	Writing
Band 5.0	88	89	93
Band 4.5	76.4	79.5	84
Band 4.0	62.7	59.8	63
Band 3.5	39.7	41	36

* Based on total group of examinees tested from August 1987 through to January 1992 (n=3,181), test population predominantly Chinese.

Table 2 Intercorrelations Among Test Totals (N=52)

Test	1	2	3	4
1. CanTEST (Time 1)	–			
2. TOEFL (Time 1)	.74	–		
3. CanTEST (Time 2)	.85	.84	–	
4. TOEFL International (Time 2)	.84	.76	.77	–
5. TOEFL Institutional (Time 2)	.71	.71	.71	.76

Table 3 Cross Tabulation of TOEFL Score by CanTEST Band Score
Canada Indonesia Language Program (N = 51)

		CanTEST Bands						
		3.50	3.75	4.0	4.25	4.50	4.75	5.0
TOEFL SCORES	-469	1						
	490-509	1	1	1		2		
	510-529		1	3	5	4	1	
	530-549			1	3	4	4	
	550-569				3	1	5	3
	570-589						1	3
	590-609						1	2

Chi-Square	Value	DF	Significance
Pearson	71.70	36	.0003
Cramer's V	.484		.0003

Figure 1a **Contingency Table for CIPP (Jakarta)**

CANTEST \ TOEFL	PASS	FAIL
PASS	14	1
FAIL	7	30

Decision Agreement = 85%
N = 53
High Stakes: CanTEST yes
TOEFL yes
TOEFL Cut Score = 550
CanTEST Cut Score = Band 4.5

Figure 1b **Contingency Table for Overseas Training Office (Jakarta)**

CANTEST \ TOEFL	PASS	FAIL
PASS	13	22
FAIL	24	70

Decision Agreement = 64%
N = 129
High Stakes: CanTEST no
TOEFL yes
TOEFL Cut Score = 550
CanTEST Cut Score = Band 4.5

Figure 1c **Contingency Table for CCLC (Beijing)**

CANTEST \ TOEFL	PASS	FAIL
PASS	13	9
FAIL	4	23

Decision Agreement = 74%
N = 49
High Stakes CanTEST yes
TOEFL no
TOEFL Cut Score = 510
CanTEST Cut Score = 4.0

Figure 1d **Contingency Table for February-March '93 TOEFL**

FEB TOEFL \ MARCH TOEFL	PASS	FAIL
PASS	9	7
FAIL	4	19

Decision Consistency = 72%
N = 39
High Stakes: February yes
March yes
TOEFL Cut Score = 550