DOCUMENT RESUME

ED 055 484 FL 002 527

AUTHOR Blatchford, Charles H.
TITLE A Theoretical Contribution to ESL Diagnostic Test Construction.
PUB DATE 7 Mar 71
NOTE 12p.; Paper presented at the Fifth Annual TESOL Convention, New Orleans, La., March 7, 1071

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Criterion Referenced Tests; *Diagnostic Tests; Educational Experiments; *English (Second Language); Norm Referenced Tests; Statistical Analysis; Statistical Data; *Tables (Data); Test Construction; Testing; Test Interpretation; *Test Reliability; Test Results

ABSTRACT

A diagnostic test in English as a second language should be a series of miniature tests on specific problems. Subscores in each area should be considered rather than a total score. The results should be used to probe mastery in an area rather than provide the means for comparing one student against another. The statistical reliability of the results does not necessarily depend on test length. The teacher should look at each item for each student rather than the score and should spend more time studying the analysis of each student's test. The criterion of the percent of correct decisions may be a more meaningful measure than ascertaining the traditional coefficients of reliability. Tables provide the statistical data under consideration. (VM)

ABSTRACT

A Theoretical Contribution to ESL Diagnostic Test Construction

Charles H. Blatchford
University of Hawaii

This paper considers the results of an experimental 40-item diagnostic test dealing with 10 grammatical mistakes typically made by Chinese students; the analysis focuses on the scores of these 10 mini-tests. The purpose of the experiment was to calculate the reliability of the mini-tests and then to determine how many items are needed to establish "good" reliability.

Two forms (A & B) were administered a week apart to 298 ESL students. Validity of the mini-tests was checked by constructing a composition with the same grammatical mistakes and asking the students to identify them.

Reliability coefficients (K-R #20) ranged from .67 to .91. The data were then analyzed as if each mini-test in Form A had only 3 items, and then only 2 items; r ranged from .61 to .87, and from .28 to .82 respectively.

From a different point of view, the optimum number of items may be suggested by asking how much useful information is lost if a decision is made on the basis of 2 items rather than 4. If the criterion is the student's consistently good, or poor, performance from A to B, the degree of such consistent performance is very stable whether based on 4, 3, or 2 items per subtest.

"A Theoretical Contribution to ESL Diagnostic Test Construction"[1]

[1]Much of the content of this paper, which was presented at the TESOL Convention, New Orleans, March 7, 1971, is derived from my Columbia University dissertation, "Experimental Steps to Ascertain Reliability of Diagnostic Tests in English as a Second Language" (Ann Arbor: University Microfilms, 1970, Order #70-18,785).

Charles H. Blatchford
University of Hawaii

This paper is addressed to some problems in diagnostic testing, and I should probably start out by defining just what a diagnostic test is. In TESL we usually think of A.L. Davis' "Diagnostic Test for Students of English as a Second Language"[2] as a prime example of a test in this category.

[2]A.L. Davis, "Diagnostic Test for Students of English as a Second Language" (Washington: Educational Services, 1953, and now distributed by McGraw-Hill).

The difficulty is that when the test is given, it most likely loses its diagnostic character, because its score is reported as a single number.

First, then, my definition of a diagnostic test is functional, and depends on the way scores are reported: whenever several part scores are reported for a test, something more than that global concept of "English" is being tested, and certain aspects are therefore diagnosed, no matter whether the test is billed as an achievement test, a proficiency test, or whatever. In other words, the degree to which a test is diagnostic depends not so much on the purpose of the test, but on the way in which scores are analyzed. Let us consider TOEFL for a moment: TOEFL is usually considered to be a proficiency test, and when its total score is considered by an admissions officer, it can quite rightly be so classified. However, if one looks at the five part-scores for reading comprehension, vocabulary, and so

on, the test is serving a diagnostic purpose, in that information about an individual's particular strengths and/or weaknesses is obtained. That is, we have specific information not on "English," but on certain abilities or skills.

Second, my definition of the ideal diagnostic test is that it be criterion-referenced, not norm-referenced. That is to say, one should look at whether mastery of the content has taken place--comparison with a criterion--rather than at how a student fares in relation to others--comparison with a norm. Although I just cited TOEFL as one example of gross diagnosis, it is a norm-referenced test, and the scores will not help inform the classroom teacher about specific weaknesses. The Davis test, on the other hand, is a criterion-referenced test. But *unless* the answer sheet is very carefully studied, the test with its one score will not give the teacher much information on strength or weakness. Usually, it is used as a placement test since its score is translated into specifications of how much more English a student should study. To summarize, first, a diagnostic test should have subscores; and second, it should not even have a total score, so that the temptation to make norms will be avoided.

In essence, a diagnostic test should be considered as a series of miniature tests on specific problems. But as soon as one considers short tests, there is the difficulty of statistical reliability--that index of how stable an individual's performance is from one form of a test to another. Reliability is felt to be dependent on test length: the longer the test, the more reliable. But, with many tests, we cannot afford great length. As Thorndike and Hagen put it, "Diagnostic testing faces a very troublesome dilemma. How is the test to provide sufficient diagnostic detail,

and yet appraise each separate ability with sufficient reliability?"[3]

[3]Robert L. Thorndike and Elizabeth Hagen, *Measurement and Evaluation in Psychology and Education* (New York: Wylie, 1961), p. 297.

To attack this problem of the reliability of miniature tests, an experimental, untimed, 40-item instrument was constructed to test ten grammatical problems, not general abilities. Examples of such problems are the use of wish and the patterns its use requires; if and "contrary-to fact" conditions; the use of because and therefore as connectives; the use of since, for, and ago; and so on. Each of these ten grammatical problems was tested by four multiple-choice items and the options were based upon Chinese students' mistakes. For example, two of the four items testing wish were as follows:

I can never finish my work. I wish I (1) have more time.
(2) to have more time.
(3) could have more time.
(4) have had more time
(9) I don't know the answer.

It takes an hour to get to school.
I wish I (1) could live nearer.
(2) have lived nearer.
(3) to live nearer.
(4) live nearer.
(9) I don't know the answer.

Two of those testing for, since, and ago were as follows:

I have been watching TV (1) for an hour.
(2) since an hour.
(3) an hour ago.
(4) from an hour.
(9) I don't know the answer.

I have been living at 350 Main Street (1) two years ago.
(2) from two years.
(3) for two years.
(4) since two years.
(9) I don't know the answer.

It can be seen that the items are structurally similar, although the options are given in different (randomized) order.

To 298 secondary and college foreign students, two forms of the test were administered a week apart, so that a Pearson product-moment reliability measure could be made. For each of the ten grammatical problems, there was then a reliability coefficient. Such product-moment reliability ranged from .37 (#2) to .79 (#6) as seen in Table 1.

Table 1 about here

By Kuder-Richardson Formula 20 for internal consistency, the ten coefficients ranged from .67 (#9) to .91 (#6). "Good" reliability is considered in the .90's or high .80's.[4]

[4]David P. Harris, *Testing English as a Second Language* (New York: McGraw-Hill, 1969), pp. 16-17.

Table 2 about here

The reliability figures were then recalculated on the miniature tests by dropping one of the four items and thus considering each mini-test as having only three items. Each reliability figure drops. Similarly, when each mini-test was considered to have only two items, the coefficients dropped yet again. The range of these coefficients was from .28 (#9) to .82 (#6). Still, in many of these mini-tests, there is good internal consistency reliability, or at least it can be considered to be good, when there are, after all, only two items making up each test!

It may now be asked what these data say regarding the optimal number of items per mini-test. It seems that for most purposes, where one is interested in descriptions of, rather than decisions about, individuals, a test of two items per problem tested may be sufficient.

From another point of view, the question of reliability can be considered not in terms of either internal consistency or product-moment coefficients. The question of how long the test should be may be rephrased to ask how much useful information is lost if a diagnosis of a student's English is based on a mini-test of two items rather than four. To attack this problem, let's look at a hypothetical situation. Four correct responses out of four will be classified as [+] and 3, 2, 1, or 0 right as [-]. For example, if on Form A a student gets two items out of four right, the student will be classified as [-] by this criterion. Should the teacher decide to teach him another lesson on the given problem? Let's say a decision to teach is made. If on Form B (given a week later but with no intervening instruction) the student scores two out of four again (classified as [-]), the correct decision was made. His performance was consistent in a negative way [-,-]. Conversely, if a student got a score of four on Form A (classified as [+]), and a four on Form B [+], and if the decision not to teach more had been made, the consistency of his performance [+,+] also corroborates the decision as being right, this time in a positive way. Thus, similarity of performance, [+,+] or [-,-] is the basis for determining whether the correct decision has been made.

Let us look at some of the data in this light. The first line in Table 3 can be read as follows: 66 students who got four right on Form A got four right on Form B; 127 who got less than four right on Form A got less than four right on Form B. The students classified in these two cells,

[+,+] and [-,-] performed consistently from one testing to the next, and for them a correct decision was made, that is, the [+,+] cell members needed no further instruction, and the [-,-] cell members did. Correct decisions were made for 193 cases, which are .647 of the total of 298.

Table 3 about here

Thus, if one had based his decisions just on Form A performance, his decision would have been corroborated in 65% of the cases. Or,put another way, assessments of a student's knowledge based on Form A performance seem to be borne out against the criterion of Form B performance in 65 out of 100 cases. The numbers in the other two cells indicate erroneous assessment. Thirteen students who got less than four right [-] on Form A performed perfectly on Form B, and 92 who performed perfectly [+] on Form A got less than four right [-] or Form B. Their inconsistent performance would have led to mistaken assessment and placement. In mini-tests one through ten, the percentages of correct assessment range from 62% (#2) to 79% (#6). If one decided from chance alone, or if one had no prior knowledge of the examinees, one would expect to be right 50% of the time. The percentages just given thus improve decision making. If one decided only on the basis of Form A, 53% (158 out of 298); if on the basis of Form B only, 27% (79 out of 298).

The figures and percentages just discussed are those for Form A when four items constitute each mini-test. When the number on Form A is reduced from four to three (as shown in the next column of Table 3), the percentage of examinees performing consistently declines, but only very slightly. When

the number of items is further reduced to two, the percentage decreases a maximum of five percentage points from what it was when the mini-test comprised four items. And in set six, which generally appears to have the best Kuder-Richardson Formula 20 reliability, there is even a tiny gain! To summarize, when it comes to the percent of correct decisions, the shorter mini-tests seem to give as much information as the full four items. The median percent of correct decisions when the test is four items long is .69, and when it is two items long, is also .69. It appears that the additional two items do not provide much, if any, more information.

So much for the theoretical side. What about the practical? I assume that since there are not many diagnostic tests, most are made by the teacher. What does the information above mean for the teacher when he is constructing a test?

1. I believe it means that with confidence he can use only two items per problem and be fairly sure of his diagnosis.

2. I believe it means that he should look at each item for each student--not using total scores. This procedure will obviously require much more time, but unless it is followed, the time spent in testing is not really worthwhile.

3. I believe it means that he can individualize instruction to a greater extent <u>if</u> he is willing to spend more time in studying the analysis of each student's test. Such individualization will require the abandonment of set ways. It will mean that he not give his pat diagnostic test at the beginning of the term, generalize about total scores, and then proceed blithely with the set syllabus. If that procedure is followed, both criteria for a diagnostic test with which this paper was introduced are being discarded.

In conclusion, provided that test-makers follow the usual canons of carefully constructing and pre-testing items, I believe the teacher can

trust the diagnostic nature of his results even if the mini-tests on each grammatical problem contain only two items--or even only one, and if sufficient time is spent looking at the test papers, not the score. Using the criterion of the percent of correct decisions made is perhaps a more meaningful measure than ascertaining traditional coefficients of reliability.

Table 1

Product-Moment Reliability Coefficients When

Forms A and B have n Items in Each Miniature Test

(N = 298)

Mini-test	$r_{A^4B^4}$	$r_{A^3B^4}$	$r_{A^2B^4}$	$r_{A^1B^4}$	$r_{A^4B^3}$	$r_{A^4B^2}$	$r_{A^4B^1}$
1	.437	.420	.411	.361	.418	.401	.398
2	.374	.369	.363	.264	.383	.371	.292
3	.445	.435	.406	.329	.423	.381	.315
4	.601	.576	.512	.358	.595	.581	.438
5	.620	.595	.586	.503	.627	.627	.579
6	.785	.759	.761	.666	.764	.744	.680
7	.462	.470	.458	.373	.455	.323	.173
8	.616	.586	.548	.525	.556	.635	.642
9	.671	.602	.531	.587	.660	.613	.408
10	.618	.582	.596	.466	.601	.572	.523

Table 2

Kuder-Richardson Formula #20 Internal Consistency Reliability

When Forms A and B Have ^{n} Items in Each Miniature Test

(N = 298)

Mini-test	A^4	A^3	A^2	B^4	B^3	B^2
1	.873	.835	.780	.875	.832	.776
2	.854	.798	.642	.726	.720	.628
3	.786	.769	.654	.778	.732	.662
4	.829	.750	.620	.797	.723	.574
5	.862	.802	.754	.689	.696	.740
6	.906	.870	.818	.909	.876	.774
7	.794	.721	.590	.615	.534	.290
8	.840	.777	.686	.685	.580	.680
9	.670	.609	.276	.704	.583	.222
10	.781	.705	.744	.848	.841	.774

Table 3

Consistency of Performance from Form A to Form B as Measured by Numbers and Percents of Examinees Getting Specified Scores

(N = 298)

Mini-test	Form B Score	Number of Items in Form A Sets: 4 # Right 0-3 (-)	4 # Right 4 (+)	%[a]	3 # Right 0-2 (-)	3 # Right 3 (+)	%	2 # Right 0-1 (-)	2 # Right 2 (+)	%	1 # Right 0 (-)	1 # Right 1 (+)	%
1	4 +	13	66	.65	12	67	.64	11	68	.63	7	72	.58
	0-3 -	127	92		123	96		119	100		101	118	
2	4 +	82	76	.62	79	79	.63	77	81	.62	63	95	.60
	0-3 -	110	30		110	30		104	36		83	57	
3	4 +	43	127	.69	33	137	.69	29	141	.68	21	149	.65
	0-3 -	77	51		67	61		61	67		46	82	
4	4 +	26	84	.74	24	86	.74	20	90	.71	7	103	.56
	0-3 -	137	51		134	54		122	66		64	124	
5	4 +	19	57	.78	18	58	.77	15	61	.75	13	63	.71
	0-3 -	176	46		171	51		163	59		147	75	
6	4 +	32	109	.79	31	110	.78	19	122	.80	14	127	.77
	0-3 -	127	30		122	35		115	42		101	56	
7	4 +	24	72	.69	21	75	.69	16	80	.64	4	92	.51
	0-3 -	134	68		129	73		112	90		61	141	
8	4 +	18	153	.64	14	157	.64	11	160	.64	6	165	.63
	0-3 -	41	86		35	92		30	97		23	104	
9	4 +	75	172	.71	72	175	.73	66	181	.73	5	242	.86
	0-3 -	41	10		41	10		37	14		15	36	
10	4 +	63	52	.70	61	54	.69	29	86	.71	24	91	.64
	0-3 -	157	26		151	32		126	57		100	83	

[a]% is the sum of the [-,-] and [+,+] cells divided by N.