DOCUMENT RESUME

ED 385 140 FL 023 132

AUTHOR Griffee, Dale T.
TITLE Classroom Testing for Teachers Who Hate Testing: Criterion-Referenced Test Construction and Evaluation.
PUB DATE 95
NOTE 20p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Foreign Countries; Higher Education; *Item Analysis; Language Tests; Norm Referenced Tests; Statistical Analysis; *Student Evaluation; *Test Construction; Testing; Weighted Scores
IDENTIFIERS Japan

ABSTRACT

This paper introduces criterion-referenced tests (CRTs), compares them with norm-referenced tests (NRTs), discusses how they can be evaluated and revised, and presents a study of an actual class and textbook test evaluation using CRTs. NRTs have dominated testing methodology since the mid-1970s; an example is the Test of English as a Foreign Language (TOEFL). CRTs are much less well known; they determine the amount of material learned rather than spreading students out along a continuum of general ability. In foreign language learning, NRTs measure general language proficiency; CRTs measure specific objectives. NRTs are of little help in diagnosing students' strong and weak points, assessing achievement, or evaluating programs. CRTs, which can be designed and evaluated by using item analysis, serve much better in these areas. The test used in the study was designed by a teacher with many years experience in teaching English as a Second Language. Unfortunately, the test described and shown, using CRT with item analysis, is found to be ineffective. Specifically, the test lacked institutional goals, forcing reliance on the textbook, rather than course objectives, for test construction. (Contains 13 references.) (NAV)

Classroom Testing for Teachers who Hate Testing: Criterion-Referenced Test Construction and Evaluation

Dale T. Griffee

Seigakuin University

Despite the increasing number of teachers who have master's degrees in teaching English as a second language (TESOL) or other forms of training, there remains an ignorance and even an aversion to the technical aspects of test construction on the part of many teachers. In the past three years, I have attempted classroom tests by several means, all of which proved unsatisfactory. Paper tests were either too easy or too difficult for my classes and interview tests proved exhausting for me. Eliminating tests altogether and basing grades on class participation and attendance also proved unsatisfactory for several reasons. First, I had the feeling of not being fair to my students. I flunked one student who was on the borderline of allowable absences, seldom participated in discussions, and on occasion slept in class. On the other hand, I gave high grades to students with good attendance but low class participation. In both cases, I felt on shaky ground and wished for additional criteria. Second, students expect a test, and I wonder how seriously they take a course without a final examination. Third, by not giving tests, I was not receiving any feedback on student progress. Fourth, without a pretest, I had no idea what the level of my students was on entering my class or what their level of previous knowledge was. Fifth, without tests, especially a final test, I was not allowing a sense of closure and completion to my course. The purpose of this paper is to introduce criterion-referenced tests to teachers who either dislike the whole idea of testing or who have for one reason or another avoided the issue of testing. Two questions will be addressed: What is the difference between criterion-referenced tests and norm-referenced tests? And, how can criterion-referenced tests be evaluated and revised?

Norm-Referenced Tests and Criterion-Referenced Tests Defined

Most classroom teachers are familiar with norm-referenced tests (NRTs) by function if not by name. One of the well-known NRTs is the Test of English as a Foreign Language (TOEFL). However, few ESL/EFL teachers are familiar with the concept of criterion-referenced tests (CRTs), perhaps because CRTs have not been discussed as much in the literature as NRTs. For example, in explaining NRTs and CRTs, one popular teacher training text (Savignon, 1983, p. 240) gives four paragraphs consisting of fifty-one lines and two tables to explaining NRTs, but gives only one paragraph consisting of six lines to explaining CRTs. Perhaps unfamiliarity with CRTs is due to the fact that NRTs have dominated testing since the mid-1970's (Bachman, 1989, p. 248). Perhaps another reason NRTs are more familiar to teachers than CRTs is because NRTs are used to decide proficiency and placement issues which are of high interest to both program administrators and classroom teachers. For whatever reason, the distinction between NRTs and CRTs is only recently being recognized by TESOL teachers (Brindley, 1989, p. 49; Brown, 1990a, p. 125; Brown, 1992). Table 1 summarizes the differences between NRTs and CRTs.

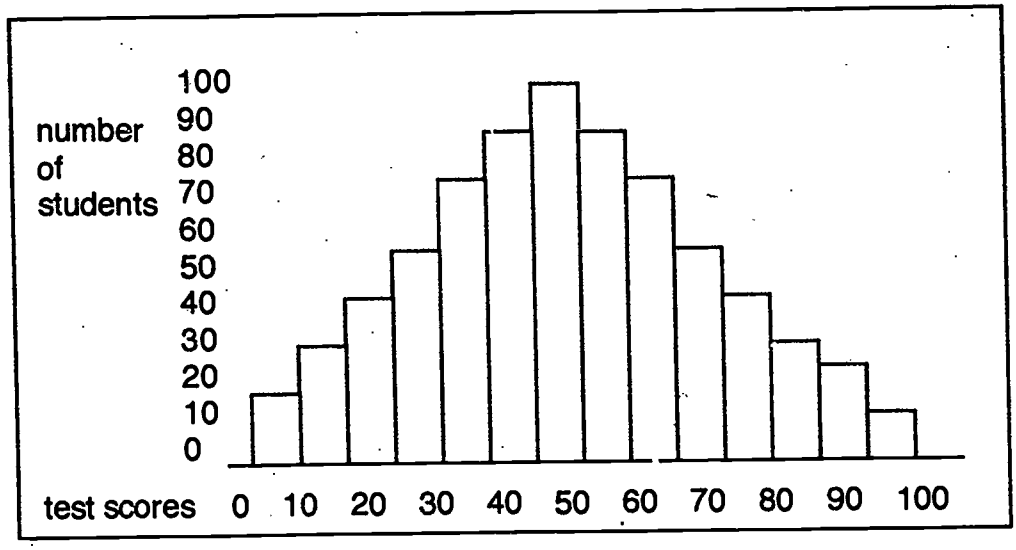
Table 1. Differences Between Norm-Referenced and Criterion-Referenced Tests Adapted from Brown (1989).

Characteristic	Norm-referenced	Criterion-referenced
Purpose of testing	To spread students out along a continuum of general ability	To determine the amount of material learned
Type of measurement	General language abilities are measured	Specific language points are measured
Type of Interpretation	Relative: A student's	Absolute:

	performance is compared with that of all other students	performance is compared only with a pre-specified learning objective
Knowledge of questions	Students have little or no idea of what content to expect in the test. questions	Students know exactly what content to expect in the test questions
Score distribution	Normal distribution of scores, e.g. a bell curve	If all students know all the material, all should score 100%

NRTs measure general language proficiency whereas CRTs measure specific objectives (Brown and Pennington, 1991, p. 7). An NRT cannot give specific information relative to what a student can or cannot do with language, and it is this characteristic that leads Bachman (1989, p. 243) to say that NRTs are a poor choice for program evaluation. NRTs interpret student scores relative to other student scores, whereas on CRTs students' scores are interpreted relative to an absolute standard, e.g., learning 25 vocabulary words by the end of the week. For NRTs to successfully compare students, they must involve a large enough sample of students to create what is called a normal distribution (see Richards, Platt, and Platt, 1992, p. 249). Figure 1 shows an example of a normal distribution.

Figure 1. Normal Distribution



If you were to connect the top of each bar with a line, you would see the familiar bell curve shape. A CRT, on the other hand, does not operate on the concept of normal distribution. In fact, a good CRT would have a positively-skewed distribution for the pretest, as seen in Figure 2, and a negatively-skewed distribution for the posttest as seen in Figure 3.

Figure 2. Positively-skewed distribution

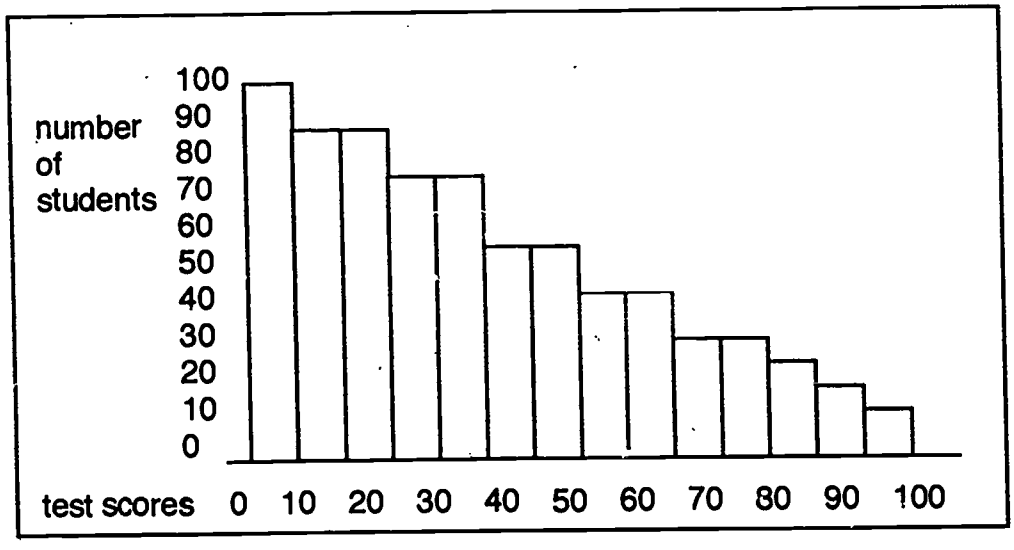
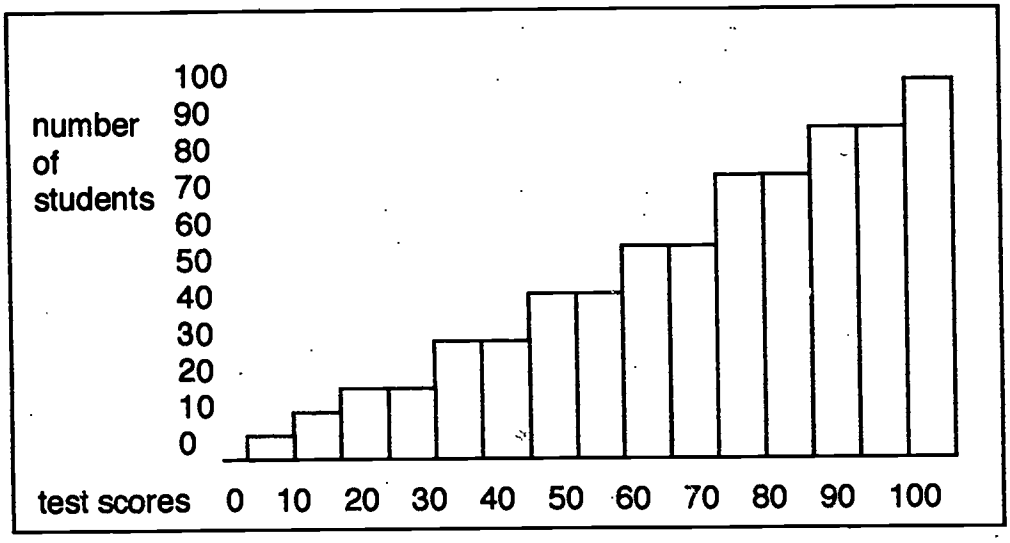


Figure 3.

Negatively-skewed distribution



The reason a well functioning CRT pretest might produce the distribution in Figure 2 is that at the beginning of a course it might reasonably produce many scores at the low end because students did not have the knowledge or skill being tested. In other words, in an ideal situation a CRT pretest would indicate that many students did not know the material and scored zero or close to it. At the end of the course, when the students had the benefit of instruction and took the posttest, they would probably score very high, which would result in a distribution like that shown in Figure 3.

Method

Subjects

The subjects in this study were 50 second-year students at a private Japanese University in a newly-formed Division of Euro-American Studies. The students were in two intact classes, one of which met on Tuesday and another which met on Thursday. Each class met once a week for 90 minutes. The Tuesday class had 25 students consisting of 11 women and 14 men, and the Thursday class had 25 students consisting of 12 women and 13 men. The students ranged in age from 19 to 21 years of age. Due to absences and late registration, 43 students (21

women and 22 men) took the pretest and 37 students (20 women and 17 men) took the posttest. All students were Japanese and all except two were from Saitama prefecture and the nearby Tokyo area. No university or department objectives existed at the time of this study and no TOEFL or other NRT test scores were available. The course syllabus was entirely up to the instructor. One grade was to given to each student for the entire year based on attendance, homework, and the final test.

Materials

In light of the absence of any institutional course objectives, the test was based entirely on the course textbook More HearSay (Griffee, 1992). Each unit in the textbook has approximately an equal number of listening and speaking exercises. The general criterion for test construction was that the test reflect the course book as much as possible. Other more specific criteria were that the test be a written test, at least half listening, that there be fifty items scored two points each, that the test contain no material directly taken from units one and two, that all questions be scorable as right or wrong, that all content items be explicitly taught in the text, and that, except for units one and two, the test items cover as much of the text as possible. Eleven formats used in the textbook were identified, and five were judged acceptable for inclusion in the test: cloze passages, multiple-choice questions, listen and write the word/number/prices you hear, listen and circle the word you hear, listen and identify what is described, and write the phrase you hear. Fourteen content areas were identified and eight were used because they had a wide distribution throughout the textbook: vocabulary, cultural items, numbers, schedules, cities, money, food, and travel. The final form of the test, not given here for test security reasons, had nine sections with a total of fifty items. The first six sections of the text, consisting of 26 questions, were for listening and included the following subtests: listen and write what you hear, count the number of words you

hear in these sentences, listen and identify which state in the U.S. is being described. The last three sections contained only written items, which included matching words and specific content questions such as "What does ASAP mean?"

Procedures

The pretest was administered in April 1993 during the second class meeting of both classes. In both classes, the first meeting was taken up with a course introduction and Unit 1 of the textbook. No pretest makeup was administered to any student who was absent or transferred into the course after the second meeting. The posttest was administered in January 1994 during the last class session. Both the pretest and the posttest were administered using the same cassette tape, which included instructions as well as the listening passages. The tests were then collected and graded by the instructor, but not returned to the students. The statistics were calculated on a Macintosh LC 520 using the Claris Works version 2.0 spreadsheet program.

Analysis

In this paper three types of statistics will be discussed: descriptive statistics, item statistics, and consistency estimates.

Descriptive Statistics

Descriptive statistics, as the name implies, give a basic description of the test. In this paper, seven descriptive statistics will be given. Five of the statistics are self-explanatory: they are the number of students, the number of test items, the minimum score, the maximum score, and the range of scores. The remaining two statistics, the mean (which is symbolized by the letter M and standard deviation (SD), require some explanation. According to Richards, Platt, and Platt (1992, p. 349), the mean is the average of a set of scores. In other words, the mean is the sum of scores divided by the number of test scores. If the scores on a certain test are 2, 4, 6, and 4 their total is 16. Divide this sum by four (the number of scores) and

the mean is 4. The standard deviation is an average of the difference of each score from the mean. The word "deviation" refers to how far each score is from the mean and the word "standard" is a kind of average. The formula for the standard deviation is as follows:

$$SD = \sqrt{\frac{S\ (x - M)^2}{N}}$$

Where: SD = standard deviation
M = mean
x = scores
N = number of students
S = sum

Item statistics

In a language test, each scorable piece of language is called an item, and a test question may contain one or more items. Item analysis is a way of obtaining some simple statistics to analyze the items on the test. Item analysis might be a new concept for many teachers, but it should be interesting for classroom teachers working with tests because it gives them a practical tool to evaluate, revise, and improve their classroom tests. Item analysis tells the teacher how students scored on each item. By using item analysis, a teacher can determine how well or how poorly each item in the test is functioning. The teacher can then revise the test by deciding which items to leave in the test and which items to delete or change. Two item statistics will be discussed later in this paper. They are item facility (IF) and the difference index (DI).

Consistency estimates

Consistency or dependability estimates for CRTs are comparable to the NRT notion of reliability. The central issue is the degree to which the teacher can expect the test to give the same results test after test. Because the main focus of this paper is on item analysis, only one consistency estimate will be given here, the Kuder-

Richardson formula number twenty-one (KR-21). KR-21 will be explained in more detail later.

Results

Table 2 Descriptive Statistics

Test	N	total possible	M	SD	Min	Max	Range
pretest	43	100	52.047	13.756	26	80	54
posttest	37	100	66.590	12.577	34	96	62

The pretest results show that, of the fifty students enrolled, only forty-three took the pretest. The results also show a fairly wide spread of 54 points ranging from a low of 26 to a high of 80 points. The mean or average is about 52 points. Since the standard deviation was about 14 points and it is known that 34% of the test scores are one standard deviation plus or minus from the mean, 68% of the students scored from 38 to 66 points. Assuming the traditional pass-fail cut point at .70, that means 84% of the students failed on the pretest. This indicates that the test was effective in that the majority of students did not know the material when they entered the class at the beginning of the school year.

The posttest results show that thirty- seven students were still in the class when the final posttest was administered. The mean score increased from 52 to almost 67 points and the minimum and maximum scores also indicate some improvement. Another way to compare pretest and posttest scores is visually through the use of bar charts. You have already seen bar charts to show the normal distribution and skewed distributions. Figure 4 shows a bar chart in the horizontal view showing the distribution of pretest scores and Figure 5 shows a bar chart in the

horizontal view showing the distribution of posttest scores. One student scored 70 on the pretest and 66 on the posttest. All other students improved from the pretest to the posttest.

Figure 4.

Bar chart showing students pretest scores.

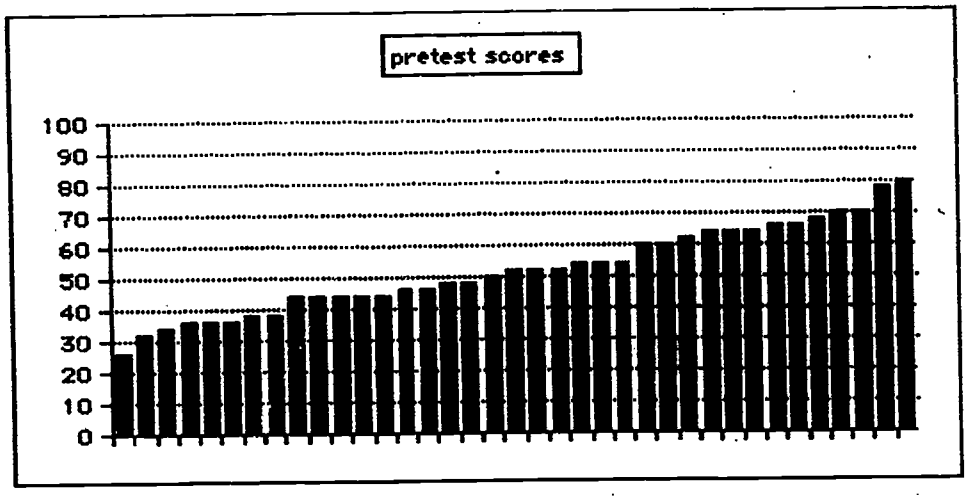
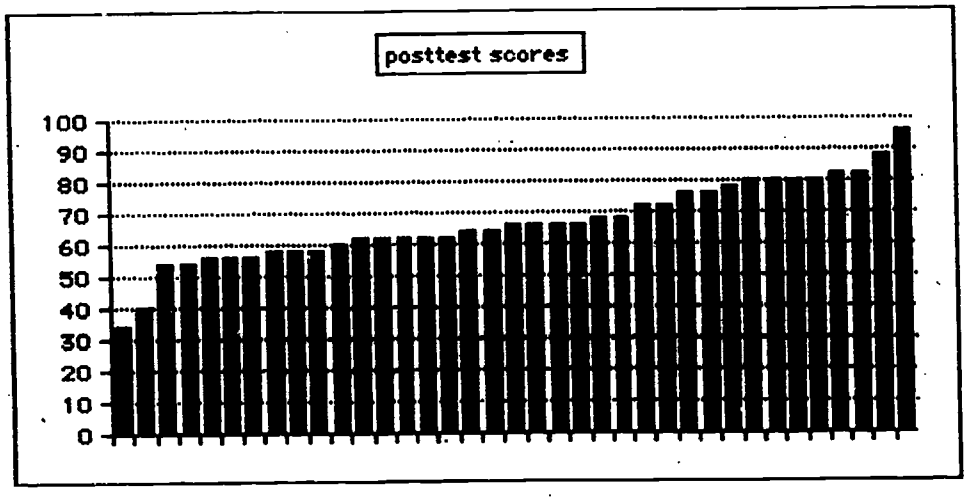


Figure 5.

Bar chart showing students and posttest scores.



In this paper, only the item analysis statistics for items 1~15 are given.

Table 3. Item Statistics for items 1- 15.

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Pretest															
IF	.70	.95	.59	.86	.70	.41	.76	.49	.73	.86	.03	.11	.79	.33	.49
Posttest															
IF	.86	.97	.70	.89	.57	.65	.86	.59	.89	.86	.03	.08	.94	.56	.70
DI	.16	.02	.11	.03	-.13	.24	.10	.10	.16	.00	.00	-.03	.15	.23	.21

Note. IF = item facility DI = Difference index

The item facility (IF) is an item statistic which gives the percent of correct answers (Brown, 1989). The formula is IF = N correct / N total. To determine the IF, divide the number of correct answers for an item by the total number of students who took the test. For example, on the pretest reported in this paper, the IF for item one was .70 and the IF for item eleven was .03. That means item one was answered correctly by 70% of the students and item eleven was answered correctly by only 3% of the students.

There are four possible outcomes for any test item in a pretest-posttest situation. These outcomes are given in Table 4.

Table 4. Possible CRT Item Outcomes

possible outcome	pretest item	posttest item	explanation
1	correct answer	incorrect answer	student forgot or was distracted
2	correct answer	correct answer	item was too easy
3	incorrect answer	incorrect answer	item was too difficult, not taught, or not reviewed
4	incorrect answer	correct answer	ideal item for a CRT

As can be seen in Table 4, in outcome one, an item is answered correctly on the pretest and then answered incorrectly on the posttest. This is a rather bizarre situation because it means that the student knew the answer on the pretest and then forgot or in some way unlearned the answer for the posttest. In such a situation, the student may have been distracted while taking the test. Several things might distract students, and the teacher should investigate such occurrences. Distraction might come from an uncomfortable room, poor lighting, or taking the test the day after an all-night party. Another reason might be that your test item was not clear to the student in that it could be interpreted in more than one way. On the pretest, the student thought that the item meant one thing and on the posttest the student thought that the item meant another thing. In outcome two of Table 4, an item is answered correctly on both the pretest and the posttest. In this case the item was too easy, probably because it was taught in previous courses. In outcome three, an item is answered incorrectly on both the pretest and the posttest. Perhaps this item was too difficult for the students to learn, or perhaps the teacher did not adequately

review the item. It could also be the case that the teacher did not actually cover the item in class. In outcome four, the item was answered incorrectly on the pretest and answered correctly on the posttest. This is the ideal case for a CRT test item. The students came to the class not knowing this point, and due to their hard work (and your good teaching) students exited the course knowing the item.

To restate, the purpose of item analysis is to provide item statistics that enable the teacher to decide which of the four possible outcomes each item belongs to, so that the teacher can improve the test by deciding in future versions of the test which items to keep, which items to revise, or which items to reject. Let's see how this process works with the first fifteen items in our test. The IF index and how it was derived has already been explained. An IF index is calculated for each item in the pretest and each item in the posttest. The pretest IF is subtracted from the posttest IF and the resulting number is the Difference Index (DI). The formula is DI = posttest IF -- pretest IF, and the higher the DI the better. The test was arranged in sections or groupings titled "tasks". The first fifteen items of the test include task 1 (items one through five), task 2 (items six through ten), and task 3 (items eleven through fifteen).

To begin the revision of task one, we see that the DI for items one through five are .16, .02, .11, .03, and -.13. These are not very impressive numbers. They indicate that item one showed an increase in scores of .16 or 16% which is not bad, but item two is only .02 or 2 percent and number five is a disaster with a minus sign indicating a net lose probably because some students experienced outcome one in Table 3. Looking at the actual test, items one through five appear as five blank lines. Students are instructed to listen and write the number they hear. The numbers the students hear on the tape are: four hundred, ninty-nine, eight thousand, thirty-two thousand five hundred, a hundred thousand and, a hundred and fifty thousand. Each number is repeated two times. The item analysis suggests

that apart from item one, we could improve the questions, especially question five. Unfortunately, item analysis does not give us any idea of what to do. To improve the questions, we must use our knowledge and imagination. My goal was to make these items more difficult (so that more students would miss them on the pretest) and more easy (so that more students would get them correct on the posttest). What I decided, in fact, was to make the items easier by making the numbers smaller and to make the items more difficult by embedding them in a sentence. Item one was changed to, "Can you help me, I have to make twenty-five copies of this report." This sentence is taken directly from one of the units.

Items six through ten are also blank lines on the test sheet with instructions in print and on tape to "listen and write the prices you hear." Students hear various prices on the tape such as three dollars and seventy-five cents for item one and eighty-five cents for item ten. Item six seems to be functioning well with a DI of .24 but item ten has a DI of zero with a pretest IF of .86 and a posttest IF of .86 indicating that no learning took place probably because the item was too easy. In the revised test, all prices were embedded in sentences. For item six, students now hear, "Is fifty-four cents enough for an airmail letter to France?" and for item ten students hear, "A ticket for the bus is fourteen fifty and you can pay the driver."

Items eleven are multiple-choice questions. The students see four words, listen to the tape, and circles the word they heard. Items eleven and twelve were not functioning well (DI of zero and minus three) while items thirteen, fourteen, and fifteen are functioning much better. I interpreted the item analysis statistics as an indication to revise items eleven and twelve. Looking closely at item eleven, we see the four options: 1) you'll, 2) I, 3) I'll, and 4) this. The utterance students heard was, "You copy and I'll collate" which is taken from one of the dialogues in the textbook. Students are selecting answer one perhaps because it is easier to hear than the correct answer, number three. My revision was to keep the utterance and change

distracter number one to a word not containing the sound /ou/ or /I/. By doing this, I reduced the distracter interference and students should find it easier to circle the correct answer. However, item analysis next year will confirm or deny my supposition. Next school semester, this test will again be administered as a pretest in April and posttest in January, and all items will be evaluated and revised as become necessary, especially items with a DI of less than .20.

Reliability is a statistical concept which is used to estimate inconsistency in a test. Imagine a perfect test. Such a test might exist in a Platonic heaven of perfect tests, but never here on earth. On earth all we have are imperfect tests all of which contain some inconsistency. We would generally like to know how reliable our imperfect test is.

This paper uses the Kuder-Richardson formula twenty-one or KR-21 which is an NRT statistic and, as such, is technically not appropriate for CRTs. However, there is an advantage to using KR-21. According to Brown (1990b), KR-21 is a conservative consistency estimate of the phi coefficient, a CRT statistic beyond the scope of this paper to expiain. However, unlike the phi coefficient, KR-21 is easy to calculate because only three numbers are necessary, and they have already been calculated and reported in the descriptive statistics. They are the number of items, the mean, and standard deviation. The formula for KR-21 is

$$\text{K-R 21} = \frac{k}{k-1}\left(1 - \frac{M(k-M)}{k s^2}\right)$$

where k = number of items

M = the mean of test items

s = standard deviation of the test items

In this calculation M and s are based on the sample posttest raw score data, and KR-21 turns out to be .85. k = 100, M = 66.59, and s = 13.35. This indicates that the

students' scores are 85 percent reliable and 15 percent unreliable.

Discussion

In this paper, the distinction between NRTs and CRTs has been clearly illustrated. NRTs are of little or no help to classroom teachers in diagnosing their students' strong and weak points, assessing achievement, or evaluating programs. I have also shown how can CRTs can be designed and evaluated by using item analysis, which makes it possible to evaluate and improve the items. The results of the item analysis reported in this paper are sobering. The test described in this paper had been carefully designed (see Griffee, 1994) by an English native speaker with a degree in TESOL and many years of teaching experience. Despite these qualifications, many of the test items were shown by item analysis to be ineffective. This indicates that test items constructed by academically qualified and experienced teachers cannot be assumed to function as intended. Item analysis operates to flag certain test items and makes it possible for the teacher to identify those items in terms of one of the four outcomes in Table 4.

One weakness of this particular test is that the lack of institutional goals forced reliance on the textbook for test construction. Using a textbook instead of course objectives as the basis for the test raises two problems: one is the narrowness of the scope and the other is limitations of sequence. The problem of narrowness scope is that the focus of the test is restricted to a single textbook. In other words, lack of institutional program learning goals forces the teacher to make the textbook an end rather than a means toward an end. The problem of limited sequence is that, in any given course, the teacher has no way of relating or supporting other courses in the curriculum. The lack of institutional or departmental objectives means there is a risk that each course in the curriculum will become an isolated island with no bridges to the other islands.

Acknowledgements

An earlier form of this paper benefited from comments by K. Anderson and D. Reid. I am indebted to W. G. Kroehler for help in setting up the spreadsheet program.

Author's address for correspondence

Dale T. Griffee

Seigakuin University

1-1 Tosaki, Ageo-shi

Saitama-ken 362

Japan

References

Bachman, L. (1989). The development and use of criterion-referenced tests of language ability in language program evaluations. In R. K. Johnson (Ed.), The second language curriculum. Cambridge: Cambridge University Press (pp. 242-258)

Brindley, G. (1989). Assessing achievement in the learner-centered curriculum. Sydney: National Centre for English Language Teaching and Research.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. TESOL Quarterly, 23 (1), 65-83.

Brown, J.D. (1990a). Where do tests fit into language programs? JALT Journal, 12 (1), 121-140.

Brown, J. D. (1990b). Short-cut estimates of criterion-referenced test consistency. Language Testing 7 (1), 77-97.

Brown, J.D. (In press). Testing in language programs. Englewood Cliffs, N. J.: Prentice Hall.

Brown, J. D. (1992). Classroom-centered language testing. TESOL Journal 1 (4), 12-15.

Brown, J. D. (1993). A comprehensive criterion-referenced language testing project. In Douglas and Chapelle (Eds.), A new decade of language testing research pp 163-184. Washington D. C.:Teachers of English to Speakers of Other Languages.

Brown, J. D. & Pennington, M. C. (1991). Developing effective evaluation systems for language programs. In M.C. Pennington (Ed.), Building better English language programs: Perspectives on evaluation in ESL (pp. 3-18). Washington D C: NAFSA.

Griffee, D. T. (1992). More HearSay: Interactive listening and speaking. Reading, Mass. Addison-Wesley.

Griffee, D. T. (1994). Criterion-referenced test construction: A preliminary report. Journal of Seigakuin University 6, 31-40.

Richards, J., Platt, J., and Platt H. (1992). Dictionary of Language Teaching & Applied Linguistics. (2nd ed.). London: Longman.

Savignon, S. J. (1993). Communicative competence: Theory and classroom practice. Reading, MA. Addison-Wesley.