DOCUMENT RESUME

ED 129 846 TM 005 466

AUTHOR	Haladyna, Thomas M.
TITLE	The Quality of Domain-Referenced Test Items.
PUB DATE	[Apr 76]
NOTE	28p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	*Criterion Referenced Tests; *Item Analysis; *Statistical Analysis
IDENTIFIERS	*Domain Referenced Tests

ABSTRACT

The objectives of this study were to first determine whether or not the empirical item analysis of domain referenced tests (DR) was justified; and second, in the event that it was, which of a set of recommended procedures was most effective for determining item quality. The analysis that followed led to the conclusion that empirical procedures were highly desirable. When these empirical procedures were applied to test data, the results indicated that four different techniques provided almost identical information: Rasch statistics, instructional sensitivity indexes, traditional statistics, and Baysian indexes. Based on these results, it would seem that any one of these four would serve adequately. (Author/RC)

ABSTRACT

The Quality of Domain-Referenced Test Items[1]

Thomas M. Haladyna

G. H. Roid

Teaching Research Division

Oregon State System of Higher Education

The objectives of this study were to first determine whether or not the empirical item analysis of domain-referenced (DR) tests was justified; and second, in the event that it was, which of a set of recommended procedures was most effective for determining item quality. The analysis that followed led to the conclusion that empirical procedures were highly desirable. When these empirical procedures were applied to test data, the results indicated that four different techniques provided almost identical information. Based on these results, it would seem that any one of these four would serve adequately.

[1] Paper presented at the meeting of the American Educational Research Association, San Francisco, April, 1976.

There has been some concern about whether or not empirical item analysis is necessa for domain-referenced (DR) tests (Millman, 1974; Hambleton, Swaminathan, Coulson, and Algina, Note 1). The objectives of this study were to (a) examine the rationales for and against the use of item analysis with DR tests and (b) in the event that item analysis is justifiable, determine which of a set of recommended procedures is most useful.

DR Tests

The nature and distinctive qualities of DR tests should be clarified before any analysis of the role of item analysis begins. This requires the defining of two concepts: namely, DR and criterion-referenced (CR) tests. An analysis of the extensive literature of (CR) testing coupled with the recent work of Hively (1974) and Millman (1974) on DR tests should reveal that at least two types of criterion-referenced tests exist: one, the traditional objective-referenced variety, for which item analysis has long since been recommended; and two, the DR test.

The CR test is primarily distinguished from others in the manner in which it is constructed, and how it is used. That is, items are written to represent instructional objectives and test results of any examinee are compared to a criterion level to determine if achievement has been satisfactory. It has been contended that CR tests exhibit little variance; and because of this, item and test statistics are so greatly attenuated that they are inapplicable (Popham and Husek, 1969). Millman (1974) points out that the conventional CR test is actually a differential assessment device (DAD) very much like the traditional norm-referenced (NR) test. The difference between the CR and NR tests is that the former allows a CR interpretation, (therefore, labeled a "CRDAD"), while any test can yield a NR interpretation.

The DR test is distinguished from the CRDAD in three respects: (a) items are created through the use of an item writing algorithm or "item form" (Hively, 1974), (b) items are randomly sampled to test forms, (c) the examinee's test score is operationally defined as an estimate of the examinee's achievement on the entire domain of items. In the first part of this study, the role of empirical item analysis is discussed in reference to DR tests. The second part is devoted to both DR tests and CRDAD's.

On the Necessity of Item Analysis

The Argument Against Item Analysis. Millman (1974) and Hambleton, *et al*., (Note 1) have asserted that item analysis for DR tests is not appropriate, and a number of reasons are provided for this position:[1]

1. Empirical item analysis leads to the selection of items based on an empirical criterion rather than on random sampling. Since random sampling is a defining characteristic of DR tests, using empirical methods for selecting items would change the nature of the test.

2. Using empirical procedures to evaluate items would result in items of moderate difficulty and high discrimination which would change the interpretation of any examinee's test score with respect to a predetermined passing standard. Selecting items empirically might eventually lead to more difficult tests which in turn would militate against the wise use of the passing standard.

3. By selecting only items that meet empirical criteria, items representing regions of the domain that were not well instructed would be omitted

[1] This rationale is presented in an abbreviated form. For a more complete discussion, consult Millman (1974, pp. 317-318 and 338-340) or Hambleton *et al*., (Note 1, pp. 17-25).

from future use. Therefore, only instructionally sensitive areas would be assessed in the DR tests, and the interpretation of test results would be different.

4. Items that measure transfer of learning may not be instructionally sensitive and, therefore, be omitted from the item pool on empirical grounds.

The Argument For Item Analysis. There are a host of reasons for the use of item analysis, many of which relate to the arguments above:

1. Empirical item analysis would be necessary to the extent that item writing algorithms are to any degree less than fully capable of being automated. An example of a fully automated item form would be one used to generate arithmetic items in which the identical words are used for each item with randomly generated numbers in the stem and/or distractors as the only variables. An example of a partially automated algorithm would be one of those suggested by Anderson (1972) involving the paraphrasing of instructional sentences and the use of word deletion or semantic transformation. In some of methods, there must be human choice as to which words to delete or the precise words used in transformations. If studies such as Roid and Haladyna (Note 2) are representative, then the science of using item writing algorithms for prose instructions would appear to be not fully developed. In that study, the use of nonautomated algorithms for prose material resulted in one item writer producing items of much greater difficulty than another item writer regardless of the method used to produce the items. Serious problems of interpretation exist when any one of these item writers contributes the preponderance of items to any DR test constructed with item algorithms that are not fully automated. Also, the problem of interpreting test results relative to

a passing standard would be complex where nonautomated item algorithms are used if empirical item analysis was not used to calibrate item difficulties.

2. Empirical item analyses would seem to be necessary to determine the acceptability of a particular item writing algorithm or item form, in any case. Words chosen for use in the item form may be confusing or too difficult for examinees or some error may be present in sentence construction or clarity of the task, and the item form would need to be revised.

3. Empirical item analysis does not destroy the random procedure so necessary in DR tests. It merely allows for the calibration of the item pool. If faulty items can be weeded out of the item pool, random selection can still occur. It should also be noted that random selection of items has long been part of traditional test theory (Lord and Novick, 1968), and is not unique to DR testing. Selecting items for a test based on empirical criteria is not a recommended practice in primary sources of measurement theory. It is a practice which evolved from the relationship between item discrimination and reliability. Selecting highly discriminating items for a test improves reliability. But this practice violates a tenet of classical test theory, that items be randomly sampled.

4. The use of transfer items may be duly noted and retained whether or not they exhibit instructional sensitivity. Empirical item analysis provides signals that items are not working. Logical analysis, by inspection of items, would still appear useful following an empirical item analysis to insure that non instructionally sensitive items are reasonably omitted.

Instructional Sensitivity. In most instructional programs, student achievement is believed to be low at the onset of instruction and high at the end (as shown in Figure 1). Performances on tests should confirm this hypothesis about achievement, and one would expect the means of instructed and noninstructed students' scores to be highly discrepant. Statistical tests of differences should confirm this large difference in test scores.

- - - - - - - - - - - - - -
Insert Figure 1 about here
- - - - - - - - - - - - - -

If the item is an analogue of the test, one might reasonably expect the same type of behavior at the item level. The difference in item difficulties of instructed and noninstructed samples given the same items is one type of CR item discrimination measure. This instructional sensitivity index comes closest to measuring, in a direct way, the effects of instruction. The most important reason for doing empirical item analysis is to reduce uncertainty about each examinee's score. Since the item discrimination index is an indicator of the degree of measurement error in that item, selecting items for the domain pool which possess high indexes would ultimately result in more reliable tests which give us more confidence in interpreting each student's test results.

A Rapprochement. The central reason for rejecting item analysis for DR tests appears to be related to a faith in item writing algorithms which proportedly produce high quality items as well as the issue of random sampling. There has not been empirical support for this belief that item writing algorithms produce fewer faulty items. In fact, what little work has been done on DR tests at the empirical level is quite negative with respect to the efficacy of the less automated, item-algorithm approaches suggested for use with prose instruction.

If faulty items occur regardless of the item writing approach, it seems reasonable to employ empirical item analysis as a first trial followed by a logical analysis of itemsas recommended by Millman (1974) and Hambleton, *et* al, (Note 1). The resultant pool of items would be of universally high quality and random selection of items to test forms can occur. Given that empirical item analysis can be fruitfully employed in DR testing, and it has long since been advocated for CR testing, which of a set of recommended procedures is most effective for creating DR and CR test items?

Item Discrimination Approaches

Techniques that may be appropriate for tests designated a CR or DR can be grouped into four major categories. Each category is discussed briefly, and several of the more prominent techniques of each category are described:

1. Instructional Sensitivity. The most prominent technique in this category is that developed by Cox and Vargas (Note 3) which is the pre-to-post difficulty index (PPDI). PPDI is computed by taking the difference between item difficulties for instructed and noninstructed groups sampled for the same item. A similar index was introduced by Brennan (1972), the notable difference being that the two groups consist of students classified as mastery and nonmastery students instead of instructed and noninstructed groups. Mastery is determined by establishing a passing standard and assigning all examinees to either category. A phi coefficient, based on the performances of one group of examinees on the same test at different times (prior to and following instruction), was introduced by Popham (1971) and studied by Tsu (Note 4) and Klein and Kosecoff (Note 5). All of these indexes have one common element: the use of two classes of student performance, one usually low achieving and the other high achieving. As noted earlier, the difference between levels of achievement prior to and following instruction for both items and tests is aptly described as "instructional sensitivity". A test showing differences in performance prior to and following instruction is sensitive to the changes in students that has actually occurred.

2. Traditional Item Discrimination. Traditional item discrimination indexes have been rejected for use with CR and DR tests due to the suspicion that variance of these test scores is so greatly restricted that these indexes could not be usefully estimated (Popham and Husek, 1969). One way to counteract the variance problem is to employ samples of instructed and noninstructed students, as in the instance where instructional sensitivity indexes are used. When the combined samples, point-biserial correlation (COMPBI) is used, the coefficients are not attenuated (Haladyna, 1974). Two other indexes that have been used in comparative studies of item discrimination include the traditional point biserial using the post-instruction only sample (POSTPBI) and the upper-lower index (computed from the difference in item difficulties of upper and lower achieving groups). The upper-lower index is a computationally simple estimate of the point biserial index.

3. Baysian Indexes. Three item discrimination indexes which are derived from Bayes Theorem were presented and studied by Helmstadter (Note 6). All Baysian indexes require collateral information in the form of preinstruction test results, or at least a lower achieving group of students. The three indexes are: (a) the probability that a student has knowledge given he gets the item right (B1), (b) the probability that a student does not have knowledge given he gets the item wrong (B2), and (c) the probability of making a correct decision--that is: high achieving or low achieving; mastery-nonmastery; preinstruction or postinstruction (B3).

4. Rasch Statistics. A fourth category of item discrimination indexes are derived from the Rasch model (Wright 1967). The Rasch model is particularly useful for studying item quality when the trait under investigation is unidimensional and the range of achievement may be greatly restricted, as is

supposedly the case with most CR and DR tests. An index of item quality is the mean square fit (MSF), a measure of the fit of item data to a theoretical item characteristics curve. A derivative Rasch statistic is formed by taking the difference between Rasch item difficulties for the same item administered to different samples. The index, called "z-difference", can serve as a type of instructional sensitivity index although there are virtually no empirical studies of its efficacy with CR or DR tests.

Empirical Studies of Item Quality

Despite the belief that conventional item quality indexes are not useful for CR and DR tests, there have not been a sufficient number of conclusive empirical studies of this issue. One reason for this may be that CR tests are difficult to construct. However, this state of affairs is rapidly changing with the increased advocacy of objective-referenced teaching and testing and item and objective banks that are conveniently packaged for teacher use. A second reason for this lack of studies might be the need for administration of these tests prior to and following instruction. The studies reviewed here are, perhaps, the most significant and contribute to the growing body of findings that describe the interrelationships among the first three categories of item-quality indexes previously discussed.

One of the earliest studies of item quality involved the comparison of PPDI with a traditional upper group-lower group discrimination index on two 40-item tests (Cox and Vargas, Note 3). Rank order correlations among the two indexes were .37 and .40 thus leading to the conclusion that PPDI measured a trait other than that measured by the upper-lower index. In a similar study (Rahmlow, Matthews, and Jung; Note 7), PPDI's and POSTPBI's were examined. The conclusions of this study were that the traditional point biserial was not a good item discrimination index and that PPDI coupled with information about item difficulties might be a more fruitful approach to studying item quality.

Popham (1971) introduced an instructional sensitivity index, a phi coefficient computed from noting performance by the same students on pretests and posttests consisting of the same items. Phi was compared with PPDI and the upper-lower index, and a lack of consistent findings were reported for seven five-item scales. Thus, none of the procedures were found to be particularly effective. Hsu (Note 4) compared the POSTPBI, phi, and PPDI on four five-item CR test scales under the conditions of varying distributions of test scores. Some of the scales did not exhibit instructional sensitivity thus minimizing the possibility of using instructional sensitivity indexes effectively. Resulting intercorrelations among these indexes revealed a strong relationship between the point biserial and PPDI indexes. The magnitudes of these correlations ranged from .43 to .95 for the 20-item scale, and from .19 to .97 for the other 20-item scale. More importantly, the higher relationships were observed when the sample was more heterogenous with respect to the achievement being measured.

Helmstadter (Note 8) and Haladyna (1974) employed objective-based achievement tests prior to and following instruction on tests ranging in size from 37 to 59 items. Both compared PPDI with several varieties of the traditional item analysis approach. The most significant departure fvom previous methods of item quality was the combining of instructed and non-instructed samples to compute COMPBI. Prior studies and some subsequent studies have focused on the POSTPBI. Using a sample of both pretested and posttested samples , the range of test scores is not restricted and PBI is not attenuated. Correlations in these studies between the COMPBI and PPDI were substantially higher than between the POSTPBI and PPDI. Further, the POSTPBI and the COMPBI were not highly correlated.

Helmstadter (Note 6) subsequently expanded the scope of his earlier study to include three Baysian indexes. The results of the analysis of the relationships between these indexes and other approaches suggested that the OOMPBI, PPDI, and the Baysian indexes were quite highly related.

Kosecoff and Klein (Note 5) compared several instructional sensitivity indexes, including the *phi* and the Brennan index with the POSTPBI and an external sensitivity index (one which required administrations of the same test both before and after instruction). Despite the fact that POSTPBI was used, instead of COMPBI, correlations among the first three indexes ranged from .82 to .97 lending support to the notion that these indexes measure the same construct, item discrimination. The external sensitivity index was not highly related to these other three indexes, and it is difficult to ascertain what is measured by it without further investigation.

In a departure from previous methodology in studying item quality, Crehan (1974) used the PPDI, the Brennan Index, and POSTBPI to select items for objective-referenced tests. Two statistics, CR reliability and validity were conceived to study the relative contributions of tests composed of items selected using each of these methods. Results indicated that PPDI and the Brennan index provided slightly superior results, but the difference was not practically significant. A speculation offered here is that if COMPBI were employed, these results might have revealed a higher degree of effectiveness for the PBI.

The studies summarized above are limited in a number of ways. First, seldom are more than a few item discrimination approaches studied. Second, the number of test items is often too few to provide sufficient information. When correlational procedures are used, the relationships among various indexes is subject to random fluctuations due to the small number of items upon which these correlations are based. Third, samples of examinees are often too small, thus lending to the possibility that estimates of item difficulty and discrimination are unstable. Fourth, instructions may be ineffective thus nullifying the use of any instructional sensitivity index. Finally, some test data may be suspect as to being CR, DR, or even objective-referenced, although what constitutes a CR or DR test is still somewhat of a debate.

- - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - -

In Table 1 the results of many of these studies are summarized with respect to what indexes were used, the resulting correlations, the subject matter of the tests, the number of items, and the sample size employed. The following conclusions are suggested by Table 1:

1. PPDI probably comes closest to measuring instructional sensitivity. It is analogous to the performance of pretested and posttested students on tests, and it is simple to compute and interpret.

2. The Brennan index appears to be highly related to PPDI, the major difference being how the high and low achieving groups are defined. Subsequent studies should reveal that the two indexes measure essentially the same construct, instructional sensitivity.

3. _Phi_ has too many restrictive assumptions to be useful. It is also computed by artificially dichotimizing groups whose scores are interval in nature. The censoring of information by grouping students and computing a _phi_ coefficient runs against common sense in studying relationships through

correlation. Students classified as nonmastery, pretest sample, or low achieving have scores which typically range widely. Placing them in one category and treating their disparate performances categorically on subsequent test items results in a loss of information.

4. POSTPBI is clearly not the same as COMPBI. It appears that the range of scores for any sample is crucial when considering these indexes. If there is a sufficient range, then either can be used.

5. Baysian indexes appear related to certain instructional sensitivity indexes, namely PPDI and COMPBI. However, there have not been a sufficient number of studies done to explore all possible relationships.

6. Rasch statistics have not been studied extensively with respect to CR or DR testing. There is a need to investigate their potential.

What would be most useful in seeking a resolution to the problem of finding appropriate measures of item quality is a series of programmatic studies where the widest range of item quality statistics are applied to legitimate CR and DR test data consisting of a sufficient number of items and examinees. The methodology used should be correlational as in most prior studies. Further, test data in this series of studies should span the widest possible range of subject matter areas rather than be restricted to mathematics as is typically the case in earlier item analysis studies.

In the next section, the first of a series of programmatic replications is reported. A wide variety of item quality indexes representing each of the four categories of techniques previously discussed was studied.

Empirical Study of Four Classes of Item Quality Indexes

Item analysis has long been accepted with CR testing, but the problem has been that of finding appropriate procedures. Based on the conclusion that both DR and objective-referenced tests are special cases of CR tests, an empirical study of the interrelationships of item statistics for items on a CRDAD was conducted.

At least four categories of item discrimination indexes were identified that may be useful in measuring item quality for CR and DR tests. If different approaches yield measures of a unique concept, namely instructional sensitivity, it is expected that the correlations among instructional sensitivity indexes should be high while the correlations between these indexes and measures of other item traits should be low. Such a finding would establish the convergent and discriminant validity (Campbell and Fiske, 1955) of the item indexes as measures of instructional sensitivity. Therefore, the questions examined were: Which item statistics appear to measure instructional sensitivity? How are these statistics related to other item statistics?

Representatives from each of four categories of item statistics were chosen. These included:

1. Rasch Statistics. The pretest sample mean-square fit (MSFPRE), the posttest sample mean-square fit (MSFPOST), the combined samples mean-square fit (MSFCOM), and z-difference (ZDIFF, an index of the difference of difficulties of pretest and posttest samples, and possibly a measure of instructional sensitivity), the Rasch pretest difficulty index (RDIFFPRE), the Rasch posttest difficulty index (RDIFFPOST), and the combined samples difficulty index (FDIFFCOM).

2. Instructional Sensitivity Indexes. The pre-to-post difficulty index (PPDI) (Cox and Vargas, Note 3) was selected.

3. Traditional Statistics. Pretest sample point-biserial (PBIPRE), the posttest sample point-biserial (PBIPOST), the combined samples point-biserial (PBICOM), the pretest difficulty index (PREDIFF), the posttest difficulty (POSTDIFF), and the combined samples difficulty (COMDIFF).

4. Baysian Indexes. The probability of having knowledge given that the student gets the item correct (B1), the probability of not having knowledge given that the student gets the item incorrect (B2), and the probability of making a correct decision, that is, assigning the student to the knowledge or no-knowledge group (B3).

Instrumentation

A 97-item CRDAD was administered to over 250 dental students at several schools prior to and following instruction on a five-volume programmed text in dental anatomy. There are three indicators of high content validity for this test: (a) Items were linked to objectives which were linked to instruction; (b) The KR-20 estimate of reliability and internal consistency was .966; (c) Average student performance on the pretest. was 33 percent and 75 percent on the posttest indicating both effective instruction and a test sensitive to instruction.

Results

To help examine the convergent and discriminant validity of discrimination and difficulty indexes for these data, the correlations among all 17 indexes are presented in Table 2. Item discrimination indexes comprise the first 11 variables, while the last 6 are difficulty indexes. Rasch difficulty and traditional difficulty indexes should not be considered similar measures of item difficulty since the Rasch indexes are sample independent while the others are sample dependent. Thus, at least three distinctive concepts or traits of item quality were expected to be measured.

- - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - -

Instructional Sensitivity Indexes. The most significant finding for these correlations is the high degree of relationship among four item discriminations measures, ZDIFF, PBICOM, PPDI, and B1. The magnitudes of the six relationships among these four variables were .79, .81, .85, .94, .96, .96. Therefore, it would appear that all four indexes measure instructional sensitivity of the items. Scatterplots of each of these six relationships among the four variables revealed uniformly straight regressions for correlations among ZDIFF, PBICOM, and PPDI. Slightly bowed regressions were observed for all correlations involving B1. This condition would, therefore, lead to attenuated magnitudes for the linear product-moment correlation coefficients involving B1. It seems reasonable to assume then that the three lowest correlations reported (.79, .81, .85) for the four discrimination indexes are actually underestimates due to the use of linear correlation to assess a curvilinear relationship. Truer estimates of the relationship might be in the mid-.90's. Thus, the conclusion that each of the four statistics measures instructional sensitivity is well supported by these data and to an extent corroborated in studies where similar analyses were done (Haladyna, 1974; Hsu, Note 4; Helmstadter, Note 6; Helmstadter, Note 8).

Of additional interest here was the rather high relationship of POSTDIFF to these four instructional sensitivity indexes. Correlations between POSTDIFF and the four indexes were .68,.72,.62, and.91. Traditionally, posttest difficulty has been believed to be non-linearly related to item discrimination with the preponderance of highly discriminating items possessing midrange difficulty and low discriminating items being very easy or very hard. With instructional sensitivity, items having high posttest difficulties are usually the most discriminating. This poses an interesting problem for the user of CR test who is interested in knowing item quality. The four instructional sensitivity indexes require that items be administered to very high and very low achieving students, such as instructed and noninstructed groups. Posttest difficult indexes can be computed from test data administered only once to the higher achieving sample. If the relationship between posttest item difficulty and item sensitivity holds with other CR or DR test data, it would be suggested that one could estimate item quality by simply noting the difficulty of items following instruction. It would be the easy items that would typically be the most sensitive to instructions. However, there may be a reason why posttest item difficulty will not be found to be a simple measure of item quality for tests different from that of the present study. Because the majority of items in the present study were highly sensitive to instruction (average PPDI was .42), this may have caused POSTDIFF to have high correlations artificially with the four instructional sensitivity indexes. That is, the majority of items in this study had both high discrimination and high posttest difficulty. For a collection of test items of less uniform quality, POSTDIFF may not correlate with PPDI at all, particularly if the majority of items are easy on both pretest and posttest.

Point-Biserial Correlations. The lack of any strong or consistent relationship of PBIPRE and PBIPOST to any other item statistic appears to support the contention of Popham and Husek (1969) that traditional item statistics like POSTPBI are useless for these CR tests. However, the reason why POSTPBI is too inappropriate is because POSTPBI is a correlation and subject to attenuation due to a restriction in the range of test scores. This restriction often occurs in pretest and posttest samples. It does not occur when the samples are combined and instruction is effective. Therefore, a point-biserial correlation, PBICOM, can be usefully employed when the sample contains a sufficient number of pretest and posttest examinees.

Baysian Indexes. Baysian indexes require the same information as PPDI and extensive computational work. While B1 was clearly established as an instructional sensitivity index, the other two Baysian indexes were not systematically related to any other indexes in this study. Thus, the utility of B2 and B3 in studying item quality appears somewhat limited. B1 has one peculiar characteristic that bears further analysis. Although it requires the same information as PPDI and is highly related (r = .79), and the regression between the two is slightly curved; it is clear upon inspecting Table 3, that any PPDI does not lead automatically to a predetermined B1. Baysian indexes are most affected by a ceiling. When a pretest or posttest difficulty level reaches its upper limit, B1 is maximized for a constant PPDI. As shown in Table 3, a PPDI of 20 might result in B1's ranging from .46 to .91, and B2's ranging from .43 to .75, and B3's ranging from .40 to .65. These deviations are systematic, monotonic with B1 but nonmonotonic with B2 and B3, and large with respect to magnitudes. While B2 and B3 may not yield useful information about item quality, B1 may prove to be a better measure than other instructional sensitivity indexes because it is sensitive to something that PPDI is not, namely, the test ceiling. But more theoretical work coupled with additional empirical studies appears necessary before this hypothesis can be verified.

- - - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - - -

Rasch Statistics. A mean-square fit can be interpreted as a z-score although in practice, mean-square fits exceeding 2.00 are normally considered grounds for rejecting the item. This is a "rule of thumb" and not based on the rejection of a statistical hypothesis. Using the criterion of 2.00 for the mean-square fit, 15 items from the Rasch analysis of the pretest sample, 6 items from the analysis of the posttest sample, and 14 items from the analysis of the combined samples would be rejected from the item pool. An analysis of the patterns of items rejected by each type of mean-square fit revealed the pattern shown in Table 4.

- - - - - - - - - - - - - -
Insert Table 4 about here
- - - - - - - - - - - - - -

The Rasch model is purported to produce sample independent estimates of item quaiity. Clearly there is little agreement among mean-square fit indexes when computed from the three analyses used in this study, pretest and posttest and combined samples which vary with respect to the range and magnitude of the trait being measured. One may speculate that the combined samples offers the most sensitive test of item quality due to the complete range of achievement represented. In that event, the pretest samples and posttest samples do not provide stable mean-square fit estimates of item quality.

Summary

In this paper, the rationales for and against item analysis of DR tests were examined. It was concluded that item analysis is necessary under several conditions, especially when the item pool is first drafted and defective items are strongly suspected to exist. Since item generating approaches which characterize DR tests do not lead *a priori* to the production of high quality items, it was argued that item analysis can be gainfully used to investigate item quality, without compromising the restrictive assumptions behind DR testing. Item analysis has long been accepted with CR testing but the problem has been that of ascertaining which of a collection of available procedures are best. Based on the conclusion that both DR and objective-referenced tests are special cases of CR tests, the relationships among 17-item statistics including 11 item discrimination indexes were studied for a CR test (actually a CRDAD). A number of practically significant relationships were observed. Most importantly, a Rasch z-difference statistic, a combined sample point-biserial, the Cox-Vargas index (PPDI), and Baysian index (B1) were found to best measure the quality of both tests and items which represent instructional sensitivity.

Closer inspection of the data, involving scatter plots and a simulated comparison between the Bl and PPDI, revealed that the Bayesian index, Bl, provides additional information that the PPDI does not. PPDI comes closest to measuring item sensitivity in a direct and simple manner which is conceptually satisfying when one considers the nature of instructional sensitivity at the test level, i.e., the difference between pre- and post-instruction means. However, Bl may ultimately be a more sensitive measure of item quality. For items with identical PPDI's, the Bl's will be higher for those items with a higher posttest difficulty. That is, Bl tells us which items with the same PPDI's provide us with more confidence that the students have knowledge given that they got the item correct.

It is apparent that a host of statistical procedures, some of which are traditional item indexes, can lead to much of the same information about item quality given that pretest and posttest data is available for the same items, and that these data are combined during item analyses.

Table 1

Summary of Studies Where Item Analysis Indexes
Were Computed on CR Tests

Study	Indexes Compared	Correlations	Median	Subject Matter	Items	Student Sample
Cox & Vargas (Note 3)	UL vs PPDI	.37 & .40	.38	Math		
Popham (1972)	UL vs PPDI UL vs PPDI	0 to 1.00 -.37 to .17	.63 .20	Inst. Concepts Inst. Math.	5 15	100 100
Hsu (Note 4)	RPBI vs PPDI RPBI vs ϕ PPDI vs ϕ RPBI vs PPDI RPBI vs ϕ PPDI vs ϕ	.35 to .93 .00 to .78 -.07 to .95 .44 to .92 .49 to .91 .25 to .94	.71 .73 .71 .80 .84 .61	Sub-Math. Math.	20 16	
Helmstadter (Note 8)	RPBI (comb) vs RPBI (post) RPBI (comb) vs PPDI	.47 .78			59	28
Haladyna (1974),	RPBI (post) vs RPBI (pre) RPBI (post) vs PPDI RPBI (comb) vs PPDI	.46 to .74 .05 to .49 .64 to .86	.57 .31 .74	Education	38-42	189
Helmstadter (Note 6)	RPBI (post) vs RPBI (comb) RPBI (post) vs PPDI RPBI (comb) vs PPDI RPBI (comb) vs Bayes PPDI vs Bayes (post)	.62 to .09 .44 & -.37 .95 & .90 -.29 to .88 -.51 to .87	 .68 .69	Statistics Psychology	59 50	43 55
Kosecoff & Klein (Note 5)	Brennan vs RPBI (post) Brennan vs ϕ ϕ vs RPBI (post)	.82 .83 .97		Mathematics	70	115

Table 2

Intercorrelations For All Item Statistics
(Based on 97 test items)

Discrimination Indexes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. MSFPRE																
2. MSFPOST	-03															
3. MSFCOM	38	09														
4. ZDIFF	-03	-04	-53													
5. PBIPRE	-08	-01	-22	30												
6. PBIPOST	-16	-29	-30	01	-20											
7. PBICOM	-10	-10	-58	96	41	10										
8. PPDI	-06	-11	-49	94	23	09	96									
9. B1 P(K|R)	-10	01	-65	81	44	00	85	79								
10. B2 P($\bar{K}$|$\bar{R}$)	13	-05	-05	64	07	-0	58	61	18							
11. B3 P(correct decision)	17	-09	02	36	-20	10	30	38	-09	83						
Difficulty Indexes																
12. RDIFFPRE	14	-17	03	00	-28	23	-03	05	-22	24	42					
13. RDIFFPOST	-01	-03	43	46	-43	13	-46	-34	-62	-02	-02	15				
14. RDIFFCOM	14	-07	27	46	-43	17	-44	-34	-59	-01	47	39	08			
15. PREDIFF	-09	15	-18	07	47	-24	-03	-21	36	-51	-80	-50	-37	-65		
16. POSTDIFF	-11	01	-59	68	52	-06	72	62	91	05	-15	-27	-83	-51	51	
17. COMDIFF	-07	06	-46	32	50	-12	36	21	68	-27	-37	-33	-87	-32	71	88
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Note: All decimals omitted

Table 3

Comparison of PPDI and the First Baysian Index
for Three Pre to Post Item Difficulty Differences

PREDIFF	POSTDIFF	PPDI	B1	B2	B3
20	40	20	.46	.75	.79
30	50	20	.53	.68	.69
40	60	20	.60	.60	.60
50	70	20	.68	.53	.52
60	80	20	.76	.46	.44
70	90	20	.84	.43	.40
80	100	20	.91	.70	65
20	60	40	.67	75	78
30	70	40	.70	.70	.70
40	80	40	.75	67	.64
50	90	40	.81	.68	.63
60	100	40	.86	91	81
20	80	60	80	80	80
30	90	60	82	83	77
40	100	60	84	96	85

Table 4

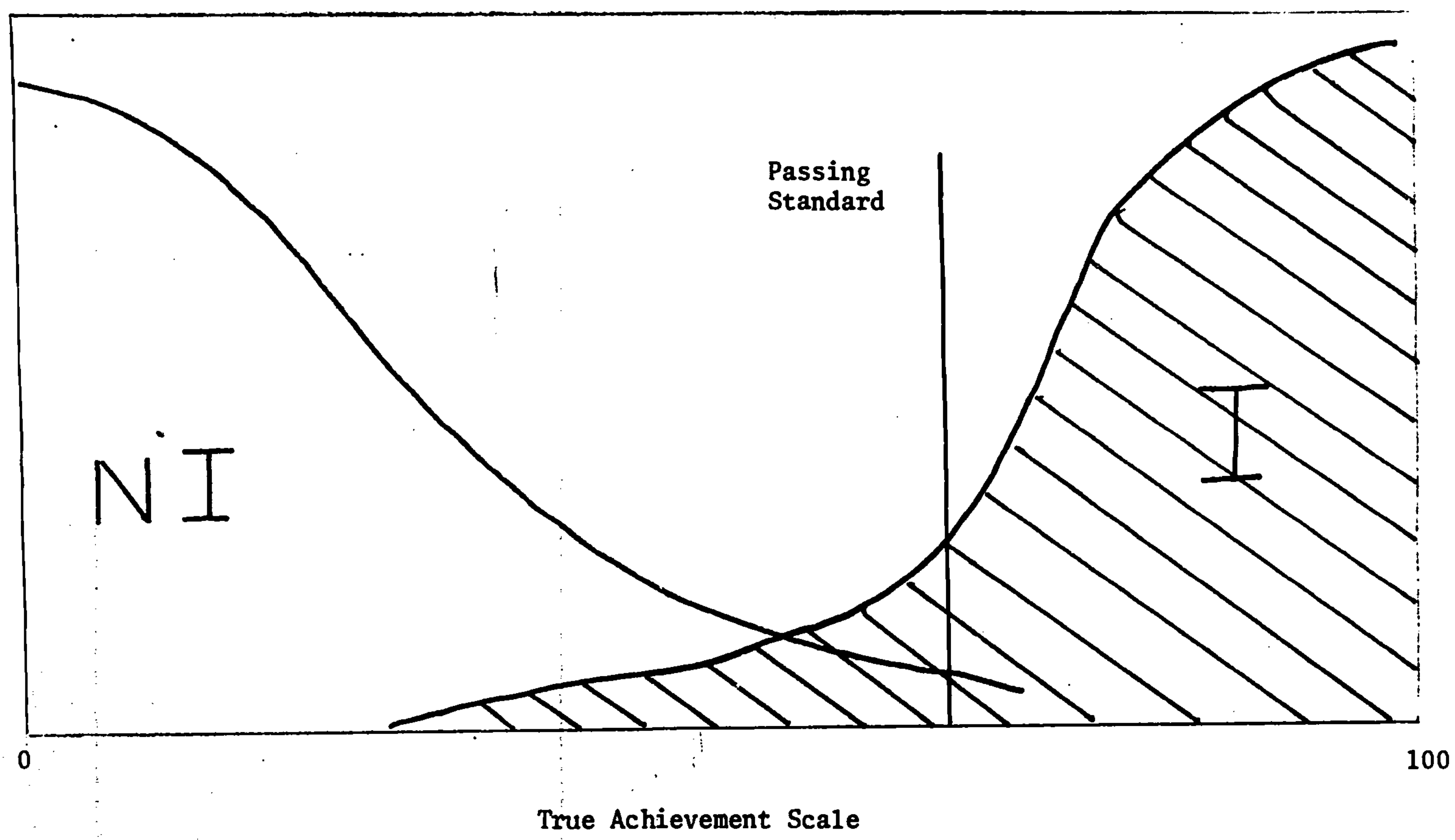
Number of Items Rejected by the Rasch Mean-Square Fit Indexes From Various Analyses

Type of Analysis	Number of Items Rejected
Pretest sample only	7
Posttest sample only	1
Combined sample.only	9
Items Rejected by More Than One Analysis	
Pretest and Posttest	1
Pretest and Combined	2
Posttest and Combined	2
All Three Analyses	2

Note: Items are "rejected" if the mean-square fit exceeds 2.00, indicating a lack of fit of the item data to an item characteristics curve.

Figure 1

Hypothetical Distributions of Instructed (I) and
Noinstructioned (N_1) Students for CR & DR tests



Note: Zero indicates absence of achievement, 100 indicates complete achievement

REFERENCE NOTES

1. Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. *Criterion-referenced testing and measurement: A review of technical issues and developments.* Symposium presented at the meeting of the American Educational Research Association, Washington, D.C., March-April, 1975.

2. Roid, G. H., & Haladyna, T. M. *A comparison of objective-based and Bormuth item-writing techniques.* Paper presented at the meeting of the American Educational Research Association, San Francisco, April, 1976.

3. Cox, R. C., & Vargas, J. *A comparison of item selection techniques for norm-referenced and criterion-referenced tests.* Paper presented at the meeting of the American Educational Research Association, 1966.

4. Hsu, T. *Empirical data on criterion-referenced tests.* A paper presented at the meeting of the American Educational Research Association, New York, 1971.

5. Kosecoff, J. B., & Klein, S. P. *Instructional sensitivity statistics appropriate for objective-based test items.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago, 1974.

6. Helmstadter, G. C. *A comparison of Baysian and traditional indexes of test item effectiveness.* A paper presented at the meeting of the National Council on Measurement in Education, Chicago, 1974.

7. Rahmlow, H. F., Matthews, J. J., & Jung, S. M. *An empirical investigation of item analysis in criterion-referenced tests.* Paper presented at the meeting of the American Educational Research Association, Minneapolis, 1970.

8. Helmstadter, G. C. *Comparison of traditional item analysis selection procedures with those recommended for tests designed to measure achievement following performance-oriented instruction.* Paper presented at the meeting of the American Psychological Association, Hawaii, 1972.

REFERENCES

Anderson, R. C. How to construct achievement tests to assess comprehension. *Review of Educational Research,* 1972, *42,* 145-170.

Bormuth, J. R. *On the theory of achievement test items.* Chicago: University of Chicago Press, 1970.

Brennan, R. L. A generalized upper-lower item discrimination index. *Educational and Psychological Measurement,* 1972, *32,* 289-303.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, *56,* 81-105.

Crehan, K. D. Item analysis for teacher-made mastery test. *Journal of Educational Measurement,* 1974, *11,* 255-262.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. *Journal of Educational Measurement,* 1974, *11,* 93-99.

Hively, W. Introduction to domain-referenced testing. *Educational Technology,* 1974, *14,* 5-10.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass., Addison-Wesley, 1968.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.). *Evaluation in education: Current application.* San Francisco: McCutchan, 1974.

Popham, W. J. Indices of adequacy for criterion-referenced tests. In W. J. Popham, (Ed). *Criterion-referenced measurement.* Englewood Clifts, NY: Educational Technology Publications, 1972.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement,* 1969, *6,* 1-9.

Thorndike, R. L. *Personnel Selection.* New Jersey: John Wiley, 1949.

Wright, B. D. Sample-free test calibration and person measurement. *Invitational Conference on Testing Problems.* Princeton, NJ: Education Testing Service, 1967.