DOCUMENT RESUME

ED 205 546 TM 810 425

AUTHOR Haladyna, Tom; Roid, Gale
TITLE A Comparison of Two Item Selection Procedures for Building Criterion-Referenced Tests.
SPONS AGENCY National Inst. of Education (ED), Washington, D.C.
PUB DATE [81]
NOTE 38p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Error of Measurement; *Latent Trait Theory; *Test Construction; Test Format; Test Reliability; *Test Theory; Test Validity
IDENTIFIERS Test Length

ABSTRACT

Two approaches to criterion-referenced test construction are compared. Classical test theory is based on the practice of random sampling from a well-defined domain of test items; latent trait theory suggests that the difficulty of the items should be matched to the achievement level of the student. In addition to these two methods of test construction, the independent variables of the study were test length and type of criterion-referenced test data, varying in sensitivity to instruction. The dependent variables of the study included two indices of the amount of measurement error present in a set of test scores. The results were consistent across four data sets. Tests created by selecting appropriate difficulty levels for students based on the Rasch model yielded smaller errors of measurement than tests which were created by randomly sampling items. This study also indicated that the relationship between measurement error and test length is a curvilinear function with the greatest decrease in error ocurring between 10 and 20-item tests. (BW)

A Comparison of Two Item Selection Procedures for Building Criterion-Referenced Tests

Tom Haladyna

Teaching Research Division
Oregon State System of Higher Education
Monmouth, Oregon 97361

and

Gale Roid

Western Psychological Services
Los Angeles, California 90025

A Comparison of Two Item Selection Procedures for Building Criterion-Referenced Tests

Within any form of systematic instruction (e.g., mastery learning), there is a need for highly relevant achievement tests to monitor achievement of individual students. Such tests have been commonly known as "criterion-referenced" (CR).

In the area of CR test reliability, two significantly distinctive conceptualizations have been discussed (Hambleton, Swaminathan, Algina & Coulson, 1978). The first refers to the consistency of correct pass or fail classifications from test to test, while the latter reflects the magnitudes of errors of measurement as it affects decisions regarding pass-fail.

Both content validity and reliability are affected by the manner in which CR tests are constructed. Essentially, test makers may develop domain specifications or objectives, create items, review these items using logical or empirical procedures, and select items for CR tests in much the manner recommended currently by test specialists (e.g., Haladyna & Roid, 1981; Hambleton, et al., 1978). The way items are selected for a CR test is an issue of major importance in CR test development and is the focus of this study.

Two Approaches to CR Test Construction

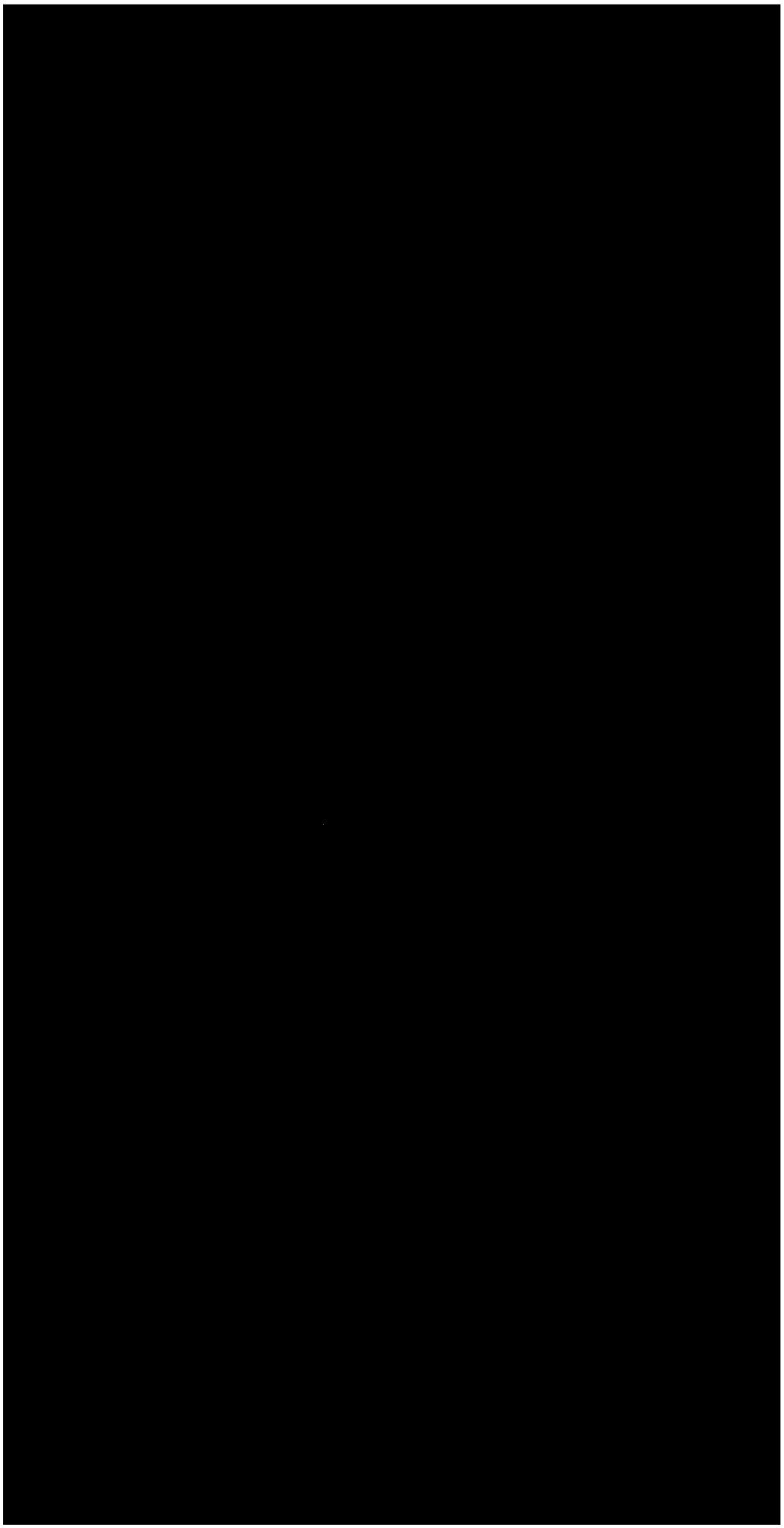
Random sampling. Classical test theory is based on the practice of random sampling from a well defined domain of test items (Lord & Novick, 1968; Nunnally, 1967). The very same approach to test construction is present in generalizability theory (Cronbach, Gleser, Nanda & Rajaratnam,

1971), and the practice of sampling is prominent in many discussions of CR testing (Brennan & Kane, 1977; Hambleton et al., 1978; Millman, 1974a, 1974b; Popham, 1978; Shoemaker, 1975).

Thus it seems desirable to randomly select items from a pool of items which have been carefully developed to represent some important instructional targets. In practice, however, we are aware that empirical procedures have been utilized via test blueprints and other means have been satisfied (Mehrens & Ebel, 1979). Most measurement textbooks give strong support to the use of the results of item analysis for selecting or removing items from achievement tests. A recent study by Haladyna and Roid (1979b) however, suggests that when item characteristic indexes are used to select items for a CR test, the results lead to larger errors of measurement when compared to tests composed by random sampling. Therefore, there is some empirical support for the practice of randomly sampling items.

Latent trait theory. Recent interest in latent trait theory has resulted in a number of research studies and applications (e.g., Hambleton, Swaminathan, Cook, Eignor & Gifford, 1978; Wright, 1977). There have been several attempts to apply the simplest of these latent trait models, the Rasch model, to CR testing (Haladyna & Roid, 1979a; Hambleton & Cook, 1977; Rentz & Rentz, 1978). In the study by Haladyna and Roid (1979a), the Rasch model seemed to be very robust in estimating student achievement despite problems with the stability of estimation of the only parameter of the model, item difficulty.

In theory, a test maker selects test items for students in such a manner that the difficulty of the items is matched to the achievement level of the student. When this is accomplished, the error of measurement for



generalizability theory and also have meaning in latent trait theory (Lord & Novick, 1968, pp. 386-387). The definitions are also generally acceptable in discussions on CR testing (Hambleton, Swaminathan, Algina & Coulson, 1978; Millman, 1974a).

1. An item universe is generated that adequately and logically represents the target of instruction, and this universe can be considered to be "unidimensional" in the sense that it represents a holistic trait.

2. A true score is the result obtained by administering all items in the item universe to an examinee in the population of examinees for which the test is intended.

3. An observed score is the result obtained by administering a subset of these items to an examinee.

4. The observed score is an estimator of the true score and is unbiased when the score is based on a random sample of items.

5. An error of measurement is the difference between a true and observed score.

It is very rare, if not nearly impossible, to obtain true scores. Yet much progress has been made in specifying content domains to the extent that finite item universes are specifiable and, in experimental conditions, entire, finite domains have been administered to samples of students (Haladyna & Roid, 1980). Thus, true scores may be directly observed. Given an item-by-person matrix of responses to items where the finite item universe has been administered, it is possible to systematically construct tests of varying length using different test construction strategies for the purpose of making comparisons in terms of errors of measurement. That

is, we can use an item-by-person matrix to construct tests using random sampling and latent trait procedures, and the simulated test results will lead to reasonable estimates of the magnitudes of errors of measurement that arise from these two approaches to test construction.

Therefore, the independent variables of the study were:

1. Two methods of test construction -- random sampling vs. selection of items based on the match between student performance level and item difficulty.

2. Four test lengths, 10, 20, 30, 40 items.

3. Four types of CR test data varying in sensitivity to instruction.

The dependent measure of the study included the absolute average deviation and a ratio of error variation and true score variation, two statistics which represent the amount of measurement error present in any set of test scores.

Sources of Data

Four item universes were administered to students prior to and following instruction. These item universes vary widely in content, educational level and sensitivity to instruction. The first two data sources contained items representing objectives which first-year dental students were to learn as part of a course in dental anatomy. The second two data sources were obtained from elementary school children as part of an instructional program assessment. All of these tests were objective-based and administered as part of instruction. Summary statistics for these CR test data are presented in Table 1. As shown there, the instructional sensitivity (pretest vs. posttest differences) of these tests varies widely, from 18.4% to 56.3%. It is also important to note that these four data sources differed in

posttest variability and levels. These four data sources seem to represent the range of situations common to instruction.

Insert Table 1 about here

Procedures

For each data source, posttest results were used, as this condition is the most prominently used in reliability and validity analyses in practice. While pretest data is desirable for other reasons, such as item analysis (Haladyna & Roid, 1981), it is expensive and difficult to obtain, and it is inefficient from the standpoint of usage of student time.

Using the person-by-item matrix for each data source, three 10-, 20-, 30-, and 40-item samples were randomly drawn from the item universe to simulate several forms of randomly composed tests of these varying lengths, a total of 12 such tests. Each of these tests were then scored using student responses to these particular items.

The Rasch model is used to support the notion that when the difficulty of a test is matched to the level of the examinee, the error of measurement is minimized. Therefore three conditions can exist when an examinee encounters a test: (a) the test is at-level and error of measurement is small, (b) the test is too difficult or too easy and the error of measurement is large, or (c) the test is near the level of the examinee and the error of measurement is moderate.

In this study, all three conditions were simulated. This was accomplished by building test forms which varied systematically in difficulty and by subdividing the sample of students into four equal quartiles. Certain combinations of test forms and student samples yielded situations where the

Table 1

Means and Standard Deviations in the Metric of Percentages Correct

for Pretest and Posttest Conditions

for Four Data Sources

Data Source	Pretest n	Pretest mean	Pretest s.d.	Posttest n	Posttest mean	Posttest s.d.	Number of Items in Domain	Sensitivity to Instruction
1	152	34.9	7.6	136	91.2	12.5	80	56.3%
2	256	32.9	13.7	254	74.8	17.9	100	41.9%
3	291	25.4	9.4	236	51.3	19.0	68	25.9%
4	306	28.5	10.7	326	46.9	15.1	68	18.4%

test was at-level, near-level, or off-level. This is illustrated in Table 2. The actual procedures used to identify the level of the test are presented in Appendix A.

Insert Table 2 about here

The consequence of this strategy was to simulate situations where students were given at-level tests where errors of measurement were predicted to be small as well as off-level tests where these errors were predicted to be the largest. Thus, two kinds of comparisons were available: (a) between the two test construction strategies, and (b) between at-level and off-level tests within the Rasch, latent trait approach.

Analysis of Data

The results of each of these 12 forms was then compared using a statistic conceived by Hambleton, Hutten, and Swaminathan (1976) for such comparisons, the Average Absolute Difference (AAD). This statistic is useful in describing the average magnitude of errors of measurement when the true scores are known. Hambleton et al., (1976) used AAD with simulated data to compare several methods of estimating true scores.

AAD is highly dependent upon the scales being used. Since random samples of items lead to percentage correct scales and the use of the Rasch model leads to an entirely different scale, a scale-free statistic, (E/T), was created which was free of this dependency upon the scale but indicated the degree of error extant in the data as a function of the true score variance. This statistic was the ratio of AAD to the standard deviation of true scores. E/T is similar to the signal-to-noise ratio discussed by Brennan and Kane (1977) except this statistic is not based

Table 2

Test Levels and Student Achievement Levels for Experiment[1]

Student Achievement Levels	Test Levels			
	Level 1	Level 2	Level 3	Level 4
1	*	01	02	03
2	01	*	01	02
3	02	01	*	01
4	01	02	01	*

[1] * Indicates appropriate level for student.

01 Indicates condition where test or student level is close to appropriate.

02 and 03 indicate a condition where student or test level is seriously too easy or too hard.

on estimates of error and true score variance but are parametric values. E/T makes possible comparisons of magnitudes of errors between tests generated by random sampling and tests created through the use of the Rasch model.

To determine the relative effects of the two test construction procedures on the dependent variable E/T, a 4 x 4 x 4 analysis of variance was done. The first variable was the method of test construction: (a) random sampling, (b) in-level, (c) near-level, and (d) out-of-level. The second variable, test length, was introduced as a control variable but also allows the study of the magnitude of errors as a function of test length. The third variable was data source. One purpose in using this factor in the design was to examine the possibility that the distribution of scores or type of distribution was a factor in explaining the degree of measurement error in these data. In a study by Haladyna and Roid (1980), when errors of classification were studied, a criterion level was determined and students were classified as pass, uncertain, or fail. The results of that study indicated that type of distribution of test scores was the major factor in determining classification errors. This third factor in the design consisted of categories of achievement, where the student sample was divided into four groups based on their true scores, the first group being the highest achieving group and the fourth being the lowest.

With each data source, there were only a small number of level tests for test lengths of 30 or 40, so interactions were not considered part of the design due to insufficient numbers of observations in some cells. The variance from these interactions were pooled with residual variance and only main effects were reported. Since the concern was for the contribution of

each main effect in explaining error variance, results were reported in proportion of variance accounted for each main effect following a test of statistical significance where alpha was set at .001.

Results and Discussion

The results of the analyses of variance for each of the four data sources are reported in Table 3. All results are reported in percent of

Insert Table 3 about here

accounted variance as all main effects were highly statistically significant ($p < .001$). Sample sizes, means, and standard deviations for all factors and data sources appear in Table 4. Of the four data sources,

Insert Table 4 about here

three proved to have sufficient conditions for the establishment of at-level tests for each test length and sample condition. For the first data set, where the sensitivity to instruction was greatest and where posttest scores were uniformly high; no level tests existed for the first three of four sample conditions studied. That is, the first three quartile groups consistently scored over 90%; and at this level, no test form proved sufficiently difficult for any of these samples to justify the designation as an at-level test. The results for the first data set are based on test scores for the fourth group only which had a wide range of achievement test scores (70 - 90%).

The results of this analysis of the sources of error variance can be classified into three categories: (a) test construction technique, (b) test

Table 3

Percent of Accounted Variance for Each Main Effect

	Data Source	Data Source 2	Data Source 3	Data Source 4
Type of Test	51.4%	13.7%	12.9%	14.8%
Test Length	40.3%	23.5%	33.1%	51.1%
Type of Sample	--	50.0%	40.8%	31.8%
Total Proportion of Accounted Variance		87.8%	96.8%	97.7%

Table 4

Sample Size, Means, and Standard Deviation for

Each Main Effect and Data Source

	Data Source 1			Data Source 2			Data Source 3			Data Source 4		
Type of Test	n	mean	s.d.	n	mean	s.d.	n	mean	s.d.	n	mean	s.d.
1. Random Sample	12	0.93	0.41	48	3.22	1.86	48	1.77	1.26	48	1.24	0.64
2. At-Level	5	0.83	0.30	29	2.54	1.59	30	1.54	1.09	18	1.18	0.73
3. Near-Level	2	1.02	0.04	30	3.49	1.69	21	2.17	1.16	22	1.84	0.98
4. Off-Level	10	1.54	0.19	33	4.59	1.91	13	3.13	1.76	24	1.82	0.61
Test Length												
1. 10 items	11	1.44	0.27	52	4.54	1.94	40	2.86	1.54	40	2.14	0.75
2. 20 items	7	1.12	0.34	32	3.38	1.69	28	1.91	1.03	28	1.45	0.48
3. 30 items	6	0.89	0.38	28	2.80	1.58	24	1.40	0.62	24	1.01	0.39
4. 40 items	5	0.75	0.49	28	2.20	1.16	20	0.8[illegible]	0.35	20	0.72	0.25
Type of Sample												
1. First Quartile				35	2.51	0.78	28	1.03	0.44	28	1.15	0.48
2. Second Quartile				35	5.28	1.90	28	2.56	1.29	28	1.82	0.78
3. Third Quartile				35	4.19	1.58	28	3.00	1.51	28	1.98	0.84
4. Fourth Quartile				35	1.87	0.75	28	1.19	0.48	28	0.94	0.40
Total	29	1.12	0.43	140	3.46	1.90	112	1.94	1.33	112	1.47	0.78

length, and (c) type of sample condition. These become the objects of further discussion.

Test Construction Approach

For the latter three data sources where the type of sample was not a problem, the approach to test construction typically accounted for a relatively small but highly statistically significant proportion of variance. In each and every data sample, the at-level tests consistently produced the smallest errors of measurement.

The criterion of effect size was used here to describe the magnitude of the differences observed. Effect size is simply the number of standard deviation units that two means differ. The differences between Rasch-based, at-level tests and randomly generated tests represented small effect sizes,[1] .23, .36, .17, and .08 respectively. While these effect sizes are small, corresponding to the proportion of accounted variance shown in Table 1, the results clearly demonstrate that when the difficulty of the tests are appropriate to the level of achievement of a particular sample, the errors of measurement are distinctly and consistently smaller.

Looking at tests that were judged to be near-level, errors of measurement were consistently higher than the at-level test results. The magnitude of these effects was .44, .50, .47, and .85. Further, these means were higher than those reported for tests where items were randomly chosen. These results should indicate the procedure for identifying level tests was valid and that near-level tests have considerably more errors of measurement than randomly generated tests as well as at-level tests. As anticipated, off-level tests were considerably error-ridden in contrast to other conditions. The one exception to this, data source four, was due to a

large amount of instability in 10-item test forms for the second and third quartiles.

The first level analysis establishes the validity of constructing achievement tests which match the level of achievement of the student. Randomly selecting items, as is advocated in classical test theory, generalizability theory, and other approaches (CR testing) where an item domain is believed to represent the object of instruction, does not produce the best tests in terms of minimizing errors of measurement. On the other hand, Rasch-based tests do. A finer level of analysis was conducted to ascertain the bias of error in estimating student scores as a function of the degree to which a test matched the achievement level of the examinees.

An examination of the AAD's (the mean difference of true and observed scores) across each condition revealed that a systematic bias did occur as a function of the difference between the level of the test and the level of the examinees. When the test form was significantly too easy, student observed scores tended to be higher than true scores. When the test form was significantly too hard, student observed scores tended to be lower than true scores.

This is a reasonable finding. The Rasch model yields domain score estimates that are higher when the group of items upon which the estimate is based are easy. Conversely, domain score estimates are deflated when the set of items is hard relative to the student's achievement.

Clearly, for high achieving students, hard tests do more harm than good. On the other hand, a student who takes an easier test is more likely to be overrated because the mismatch between a low achiever and more difficult items yield an overestimation of student achievement. In either case,

the results are larger errors of measurement which are the products of an inappropriately difficult test. The results over all four data sets show this to be consistently true.

Test Length

It was expected that errors of measurement would be greatly affected by test length. While this is theoretically predicted, the design of this study permitted a look at the magnitude of decreases in errors of measurement as a function of test length.

These results, reported in Tables 1 and 2, indicate that test length was a very significant factor, accounting for 23.4%, 33.2%, and 50.4% in three of the four data sets. In the first data set, where the distribution of scores was badly skewed, test length accounted for 40.3% of the variance. Thus it is clear that test length is a powerful factor in reducing measurement error.

The results allow us to examine the magnitude of decrease in measurement errors as a function of test lengths. These are briefly summarized below in terms of effect size.

	From 10 to 20 Items	From 20 to 30 Items	From 30 to 40 Items
Data Set 1	.74	.54	.32
Set 2	.79	.53	.32
Set 3	.72	.38	.42
Set 4	.90	.57	.38

A large effect size indicates a substantial reduction in measurement error from one test length to the next test length. From these results summarized above, it is clear that 20-item tests offer the largest increase in precision from 10-item tests and the increase between 20-item and 30-item tests is

also substantial, while the increase in precision between 30-item and 40-item tests is smallest for three of the four data sets. While it is clear that 40 item tests yield the best estimates of true scores as might be expected, 30 and 20-item tests are not that substantially inferior. In terms of overall test proficiency, these results would suggest that 20-item tests offer the most for the least, while gains made with longer tests are less substantial. Where one draws the line with respect to the number of test items is a matter of the consequences one places on making decision errors in systematic instruction (Haladyna & Roid, 1980).

Type of Sample

The third factor of the study was the type of sample (range of examinees). As noted earlier, each group of students was divided into quartiles representing four sample conditions: high, high middle, low middle and low.

Results in Table 3 would indicate that type of sample was a significant factor in determining errors. However, it must be made clear that the criterion for this analysis was the statistic E/T. As noted previously, this ratio is scale-independent. The results of Table 2 indicate that E/T is highest for the two middle quartiles where student scores varied the least.

A more useful criterion is AAD which is based on the difference between true and observed scores. While E/T is metric free, it is affected by the distribution of true scores. AAD is not metric-free but is is not affected by the distribution. Therefore, AAD was used to ascertain the amount of error extant in the data sets as a function of the four types of samples studied. Since at-level tests were the most precise in estimating student scores, these tests were studied across the three data sets where

the four sample conditions existed using a one-way analysis of variance with AAD as the dependent measure.

The results of this analysis revealed no differences as a function of sample type ($F=0.34$; $df=3.73$; $p=.80$). The means for the four respective sample conditions were: .306, .333, .338, and .343 with an overall standard deviation of .135. It was conclusive from these results that when at-level tests are employed to estimate domain scores, errors of measurement do not vary significantly with the type of sample condition.

Conclusions

Test Construction Approach

The main objective of the study was to determine if a difference existed in the magnitude of measurement errors of tests constructed two different ways. The results were consistent across four data sets which represented varying degrees of sensitivity to instruction. Tests created by selecting appropriate difficulty levels for students based on the Rasch model yielded smaller errors of measurement than tests which were created by randomly sampling items. These results offer support for the concept of latent trait theory as a basis for test construction and the practice of providing achievement tests at the functioning level of each student rather than the level of heterogeneous group of students for which a student is a member.

The results also suggest that random sampling of items is a second-best alternative, the difference between the randomly sampled tests and the Rasch-calibrated tests was not large in terms of the criterion of effect size. Nonetheless, there was a statistically significant difference in each instance.

The study also serves to show that when students receive tests that are not at their level of functioning, errors of measurement tend to be substantially higher than either randomly sampled tests and at-level tests. Thus the practice of level-testing, if the assignment of students to levels is done subjectively by human judgment, is indeed a delicate technique to employ in school assessments. When a test is appropriate to examinees, this study has served to show that domain scores are precisely estimated. When the test is not appropriate for examinees, errors are quite substantial.

The CR test developer is wise to understand the benefits and deficits of these two test construction strategies, both of which require item pools. Random sampling is a more conservative practice which guarantees a moderate but controllable amount of measurement error. Level testing provides a chance for superior precision at the expense of the chanciness when a student encounters a test that is too hard or too easy. In this respect, the Portland (Oregon) Public Schools, where such level tests are employed, uses a placement test as a form of pretest, which aims the student at the test of appropriate level. This seems to be a sensible approach, which is now grounded in research findings that support the practice.

Test Length

It is well known that test length is a powerful determinant of reliability and measurement error. This study not only provided support for this principle but indicates that errors of measurement are not evenly a function of test length. If anything, the relationship between measurement error and test length is a curvilinear function with the greatest decrease in measurement error occurring between 10 and 20-item tests and decreasing as tests reach lengths of 40 items.

As Hambleton (1979) among others has noted, one goal in CR testing is to arrive at reliable domain score estimates without unnecessarily long tests. The results of this study would suggest that test lengths of less than 20 would probably not lead to reasonable domain score estimates, but satisfactory precision can be achieved for test lengths of 20 to 30 items. Beyond 30 items, gains in precision are offset by the longer tests. This, however, is a rather subjective conclusion. One needs to set test lengths based on considerations of time allocated for testing, number of students who are likely to be classified as fail or in need of remedial instruction, and other considerations. Precision is only one of several factors that are used to determine the test length of a CR test.

It would be interesting and important to develop firmer guidelines regarding the relationship between the two. More importantly, guidelines for test length should be grounded in theory and be empirically tested to ascertain their effectiveness. How long to make a CR test is still a problem of concern.

Sample Type

It was clear for this study and from principles of latent trait theory, that errors of measurement vary as a function of the discrepancy between the student and the test. If a test is too hard to too easy, there is a bias in domain score estimation that occurs, and this bias is manifested in large errors of measurement. Despite the fact that four disparate sample conditions were employed, representing quartiles of the distribution of all examinees, no differences were found in the AAD's of these sample types. They were remarkably stable across the four sample types studied. While bias exists in domain score estimation as a result of inappropriate

level of test, it does not exist for groups of students who differ in achievement as long as the test they are given is appropriate to that level.

While this study provides strong support for the practice of building Rasch-based tests of varying degrees of difficulty to minimize errors of measurement and to achieve reliable domain score estimates, a technology for developing and using these tests in objective-based instructional programs is just emerging and requires more empirical studies which examine aspects of test construction which directly affect domain score estimation.

One of these aspects includes item analysis, particularly the stability of difficulty estimates. Haladyna and Roid (1979a) have shown that serious discrepancies in difficulty estimates obtained from different samples differ substantially, a result which Slinde and Linn (1978) observed in their study of norm-referenced tests.

In summary, this study has proven that latent trait theory, particularly the one-parameter Rasch model, has much to offer users of CR tests in precisely estimating achievement with respect to a well-defined content domain. Since domain estimation is a goal of CR measurement, the latent trait approach to CR testing holds much promise.

References

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: John Wiley, 1972.

Haladyna, T., & Roid, G. The stability of Rasch item and student achievement estimates for a criterion-referenced test. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1979. (a)

Haladyna, T., & Roid, G. Two approaches to the construction of criterion-referenced achievement tests. Unpublished manuscript, 1979. (b)

Haladyna, T., & Roid, G. An empirical comparison of strategies for decision making with criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Haladyna, T., & Roid, G. The role of instructional sensitivity in the empirical review of criterion-referenced test items. Journal of Educational Measurement, 1981, 18(1), 39-53.

Hambleton, R. K. Applications of latent trait theory to the development and use of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R. K., Hutten, L. R., & Swaminathan, H. A comparison of several methods for assessing student mastery in objective-based instructional programs. Journal of Experimental Education, 1976, 45, 57-64.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48(4), 467-510.

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

Mehrens, W. A, & Ebel, R. L. Some comments on criterion-referenced and norm-referenced achievement tests. NCME Measurement in Education, 1979, 10(1), 1-7.

Millman, J. Passing scores and test lengths for domain-referenced tests. Review of Educational Research, 1973, 43, 205-216.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, California: McCutchan Publishing Company, 1974. (a)

Millman, J. Sampling plans for domain-referenced tests. Educational Technology, 1974, 14, 17-21. (b)

Nunnally, J. Psychometric theory. New York: McGraw-Hill, 1967.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1978.

Rentz, R. R., & Rentz, C. C. Does the Rasch model really work?: A discussion for practitioners. NCME Measurement in Education, 1979, 10(2), 1-11.

Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-148.

Slinde, J. A., & Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Footnotes

This research was supported through a grant from the National Institute of Education. Opinions expressed in this paper are those of the authors and do not represent the official National Institute of Education position or policy.

[1] Effect size is the ratio of the difference in contrasted means and the standard deviation.

Appendix A

A Procedure for Assigning Tests to One of Three Categories: (a) At-level, (b) Near-level, and (c) Out-of-level

Appendix A

A Procedure for Assigning Tests to One of Three Categories: (a) At-level, (b) Near-level, and (c) Out-of-level

In this study, tests of varying lengths were systematically constructed using difficulty levels as the basis for item selection. The goal was to construct tests which varied in difficulty. Four different samples were used. Each sample was created by subdividing the population of examinees into four equal quartiles; each quartile representing a different level of achievement.

A problem remained as to identifying the appropriateness of the interaction between any test form and the level of achievement of that sample. For any sample, a test form could be appropriate to the level of examinees (+) or it could be nearly appropriate (01), or it could be inappropriate, that is too hard or too easy (02). The following procedures were developed in this study to ascertain which of the three conditions described above, +, 01, or 02 existed with each test form generated in this study.

The procedures were based on an analysis of the median and range of true scores of examinees in each quartile as well as the optimal range of test scores for a particular test. The optimal range for any test form was determined to be the range of scores for which the standard error of estimate is minimal. This range is symmetrical around the center of the scale; the size of the range was plus or minus 20 percentage points from this midpoint of the scale. For example, in a 30-item test, the optimal range was the Rasch logits equivalent to range of scores from 30% to 70% on the 30-item scale (raw score 9 to 21).

To illustrate this procedure, a 20-item test from the first data source is used. Using the fourth quartile for this analysis of the test, for the 20-item test the median was -1.34 and the range was -2.24 to -0.94. The median for the students in the fourth quartile was 1.88 and the range was 0.78 to 2.38. Obviously there was no commonality between the two respective medians and ranges, and the 20-item test form was designated 02, off-level. Where a good match between the median and optimal range of a test form and the median and range of true scores existed, the designation +, at-level, was given. When there was a close match, the designation was 01, near-level.

This procedure was applied to all four data sources to arrive at assignments of test forms. Validity of this procedure was evident in the results of the study. It was predicted that at-level tests would have appreciably lower AAD and E/T than near level and off-level tests. This prediction was confirmed in all four data sets.

The results of the application of this procedure to the four data sets are given in Tables 5, 6, 7, and 8.

Insert Tables 5, 6, 7 & 8 here

Table 5

Assignment[1] of Test Forms on the Basis of Their Median and Optimal Range for Each Test and the Median and Range of True Scores of Each Quartile

Data Source I

			First Quartile		Second Quartile		Third Quartile		Fourth Quartile	
			Median	Range	Median	Range	Median	Range	Median	Range
	Median		4.07	3.68 & 4.07	3.35	3.10 & 3.68	2.88	2.38 & 3.10	1.88	0.78 & 2.38
Test Length	for Test	Optimal Range	Assignment	Z Error	Assignment	Z Error	Assignment	Z Error	Assignment	Z Error
10	-1.71	-2.58 & -0.84	02	11.42	02	10.58	02	5.53	02	1.06
10	-1.04	-1.91 & -0.19	02	8.54	02	7.18	02	2.44	02	1.44
10	-0.39	-1.24 & 0.46	02	5.20	02	3.47	02	1.91	02	1.48
10	-0.16	-1.01 & 0.69	02	4.35	02	2.37	02	2.22	02	1.82
10	0.12	-0.73 & 0.97	02	2.80	02	1.54	02	2.87	02	1.65
10	0.41	-0.45 & 1.26	02	1.82	02	2.40	02	2.31	02	1.59
10	0.94	0.09 & 1.79	02	1.70	01	3.23	02	4.22	01	1.05
10	1.63	0.74 & 2.53	02	5.52	02	5.30	01	3.35	+	1.23
20	-1.34	-2.24 & -0.44	02	6.11	02	4.65	02	1.61	02	1.57
20	-0.25	-1.11 & 0.59	02	1.43	02	1.92	02	2.76	02	1.62
20	0.29	-0.57 & 1.14	01	2.04	02	3.08	02	3.51	01	1.00
20	1.28	0.39 & 2.18	01	7.14	01	5.97	01	2.69	+	0.70
30	-1.00	-1.93 & -0.08	02	2.39	02	1.97	02	2.28	02	1.59
30	0.00	-0.86 & 0.86	02	2.22	01	3.60	02	3.07	02	1.05
30	0.98	0.08 & 1.90	02	7.48	+	5.31	+	1.99	+	0.65
40	-0.75	-1.70 & 0.17	02	5.67	02	3.18	01	2.90	02	1.62
40	0.76	-0.15 & 1.69	01	9.49	01	5.33	+	2.02	+	0.50

[1] + indicates at-level assignment
dicates near-level assignment
dicates off-level assignment

Table 6

Assignment[1] of Test Forms on the Basis of Their Median and Optimal Range for Each Test and the Median and Range of True Scores of Each Quartile

Data Source 2

			First Quartile		Second Quartile		Third Quartile		Fourth Quartile	
	Median		Median	Range	Median	Range	Median	Range	Median	Range
			2.08	1.84 & 5.09	1.55	1.48 & 1.76	1.35	1.12 & 1.43	.89	-5.26 & 1.06
Test Length	for Test	Optimal Range	Assignment	Z Error	Assignment	Z Error	Assignment	Z Error	Assignment	Z Error
10	-2.04	-2.96 & -1.13	02	4.00	02	6.01	02	4.11	01	3.51
10	-1.07	-1.92 & -0.22	02	1.86	02	7.28	02	7.07	01	3.10
10	-0.71	-1.56 & 0.14	02	2.36	02	6.88	02	6.64	01	3.37
10	-0.40	-1.25 & 0.44	02	2.90	02	7.30	02	6.07	01	2.70
10	-0.27	-0.88 & 0.782	02	2.86	01	7.98	02	6.81	01	2.78
10	0.30	-0.55 & 1.15	02	3.27	01	6.83	01	5.79	+	2.06
10	0.54	-0.31 & 1.39	02	2.50	01	7.07	01	5.82	01	2.28
10	0.71	-0.14 & 1.56	02	4.18	+	6.03	01	5.92	01	2.41
10	1.07	0.22 & 1.92	01	3.00	+	6.49	+	5.51	01	2.71
10	1.64	0.77 & 2.51	+	2.10	+	5.49	+	4.82	01	2.18
20	-1.54	-2.47 & -0.62	02	2.18	02	6.80	02	7.29	02	3.00
20	-0.56	-1.41 & 0.30	02	3.49	02	6.89	02	5.70	01	2.17
20	0.14	-0.72 & 0.99	01	2.75	01	5.07	01	4.15	+	1.17
20	0.63	-0.22 & 1.48	01	2.25	+	4.03	+	3.34	+	1.53
20	1.35	0.48 & 2.23	+	1.98	+	3.10	+	3.00	02	1.68
30	-1.25	-2.19 & -0.34	02	2.83	02	7.60	02	6.56	01	2.41
30	-0.25	-1.11 & 0.63	02	2.80	01	4.97	01	4.05	01	1.40
30	0.40	-0.45 & 1.26	01	1.97	+	2.97	+	2.35	+	0.96
30	1.14	0.25 & 2.02	+	1.23	+	2.03	+	2.54	+	1.18
40	-1.02	-1.97 & -0.11	02	3.15	02	5.47	02	4.88	01	1.84
40	-0.24	-1.12 & 0.63	02	2.49	01	3.98	01	3.18	+	1.11
40	0.38	-0.48 & 1.24	01	1.65	+	2.21	+	1.82	+	0.86
40	0.98	0.10 & 1.87	+	0.92	+	1.76	+	1.99	+	0.98

[1] + indicates at-level assignment
icates near-level assignment
icates off-level assignment

Table 7

Assignment[1] of Test Forms on the Basis of Their Median and Optimal Range for Each Test and the Median and Range of True Scores of Each Quartile

Data Source 3

			First Quartile		Second Quartile		Third Quartile		Fourth Quartile	
			Median	Range	Median	Range	Median	Range	Median	Range
	Median		0.83	0.37 & 2.23	0.08	-0.22 & 0.37	-0.44	-0.66 & -0.22	-1.24	-3.53 & -0.66
Test Length	for Test	Optimal Range	Assignment	Z Error	Assignment	Z Error	Assignment	Z Error	Assignment	Z Error
10	-1.56	-1.56 & -0.69	02	1.64	02	5.10	02	6.04	+	1.19
10	-0.92	-1.77 & -0.07	02	1.73	02	3.94	+	3.91	+	1.34
10	-0.37	-1.22 & 0.48	02	1.55	01	3.39	+	4.21	+	1.14
10	0.09	-0.76 & 0.94	01	1.31	+	3.10	+	3.93	01	1.63
10	0.53	-0.32 & 1.39	+	0.93	+	2.82	01	4.49	01	1.65
10	0.88	-0.33 & 1.73	+	1.10	01	3.80	02	4.56	02	1.52
10	1.52	0.63 & 2.40	01	1.50	02	4.53	02	5.19	02	1.65
20	-1.24	-2.11 & -0.35	02	1.73	01	3.29	01	3.52	+	0.82
20	-0.33	-1.19 & 0.54	01	0.94	+	1.80	+	2.48	+	0.73
20	0.40	-0.46 & 1.26	+	0.60	+	1.75	+	2.45	01	1.12
20	1.18	0.30 & 2.07	+	0.78	01	2.55	01	4.01	02	1.56
30	-0.94	-1.85 & -0.44	01	1.25	01	1.89	01	2.12	+	0.64
30	-0.20	-1.08 & 0.67	+	0.65	+	1.32	+	1.42	+	0.55
30	0.93	0.03 & 1.83	+	0.56	01	1.72	01	2.76	01	1.07
40	-0.68	-1.61 & 0.24	01	0.76	+	1.15	+	1.42	+	0.43
40	0.69	-0.22 & 1.61	+	0.38	+	1.20	01	1.68	01	0.83

[1] + Indicates at-level assignment
01 Indicates near-level assignment
02 Indicates off-level assignment

Table 8

Assignment[1] of Test Forms on the Basis of Their Median and Optimal Range for Each Test and the Median and Range of True Scores of Each Quartile

Data Source 4

			First Quartile		Second Quartile		Third Quartile		Fourth Quartile	
			Median	Range	Median	Range	Median	Range	Median	Range
	Median		1.20	.88 & 2.32	.49	.06 & .86	-0.39	-0.65 & .06	-1.30	-2.66 & -.65
Test Length	for Test	Optimal Range	Assignment	# Error	Assignment	# Error	Assignment	# Error	Assignment	# Error
10	-1.40	-2.26 & -0.54	02	1.90	02	3.77	01	2.81	01	1.10
10	-0.72	-1.57 & 0.13	02	3.13	02	3.71	01	3.35	+	1.00
10	-0.37	-1.22 & 0.48	02	1.61	01	3.34	01	3.20	01	1.32
10	0.10	-0.76 & 0.94	02	1.81	+	2.24	+	3.06	02	1.72
10	0.38	-0.46 & 1.23	01	1.39	+	2.34	01	3.25	02	1.27
10	0.70	-0.15 & 1.55	01	1.47	01	3.40	01	2.90	02	1.61
10	1.46	0.58 & 2.35	+	1.13	02	3.02	02	2.93	02	1.44
20	-1.06	-1.94 & -0.19	02	1.52	02	2.47	02	2.15	+	0.68
20	-0.30	-1.16 & 0.56	02	1.27	01	1.69	+	1.47	01	0.82
20	0.31	-0.54 & 1.16	01	1.06	+	1.60	01	1.75	02	1.08
30	1.05	0.16 & 1.98	+	0.71	01	1.77	02	1.95	02	1.41
30	-0.83	-1.72 & 0.57	02	1.24	02	1.95	01	1.32	+	0.53
30	-0.18	-1.05 & 0.69	01	0.93	01	1.14	+	1.09	01	0.53
30	0.80	-0.09 & 1.71	+	0.51	+	1.20	02	1.77	02	1.07
40	-0.60	-1.51 & 0.31	01	0.76	01	1.22	+	0.93	+	0.35
40	-0.60	-0.31 & 1.52	+	0.37	+	0.91	+	1.07	02	0.81

[1] + indicates at-level assignment
01 indicates near-level assignment
02 indicates off-level assignment