ZJER

ZIMBABWE JOURNAL OF EDUCATIONAL RESEARCH

Volume 4 Number 3 November 1992 ISSN 1013-3445

Volume 4 Number 3 November 1992
ISSN 1013-3445

CONTENTS

The Application of Generalizability Theory In Constructing Achievement Tests

Donton S. J. Mkandawire,
University of Namibia

ABSTRACT

The Application of Generalizability theory to test construction provides a flexible and practical framework and yet still remains inaccessible to test developers because of its technical and mathematical approach to measurement.

Different methods of test construction to measure mastery of a universe in cognitive domain have been proposed. Simple random and stratified random sampling procedures are two such models.

The focus of this study was to examine the application of Generalizability Theory when tests have been constructed using the two models and administered to a group of students. Both coefficients alpha ($\hat{\alpha}$) and alphas ($\hat{\alpha}_S$) as indices of generalizability were computed. The findings indicated that a test constructed using simple random sampling procedure had a better coefficient of generalizability although within the comparison analysis indicated that taking stratification into account increased the generalizability of the test and that generalizability theory could be applied to the construction of achievement tests.

Introduction

In teaching, several classes of decisions are made by the teacher. The Teacher must decide the objectives of instruction prior to teaching and select the procedures and methods that will achieve these objectives (Mager, 1962). After the subject matter is taught, the student's performance is assessed in terms of instructional objectives that the teacher originally stated.

Teachers have found the task of formulating objectives rather cumbersome and difficult. However, defining objectives operationally does ease the task of later translating these objectives into domains of expected educational outputs in cognitive domain, which can then be properly assessed. Once the analysis and definition of performance in cognitive areas have been stated, then measuring the learner's attainment of objectives and determining their level of performance is easily ascertained.

A fundamental problem in the assessment of educational outputs is to determine a person's attainment with respect to a defined domain of tasks that are relevant to the desired outcomes of instruction. Since domains of tasks that define the outcomes of instruction could be infinite, tests constructed to assess attainment in these domains must necessarily be samples from the domains. A major concern in assessment should be the adequacy with which one can generalize from the sample behaviour exhibited by the examinee on a particular test, to some larger domain, or universe of tasks. A score's usefulness largely depends on its universe generalization (Shavelson et al, 1989).

Generalizability theory (GT) concerns the adequacy with which a "Universe" score can be inferred from a set of observations (Cronbach et al, 1963). The theory presents a mathematical model in the framework of which a particular test is assumed to be a random sample from a large defined domain of test items.

A candidate's score on a test constructed in this manner provides an unbiased estimate of the score on the total domain. One can, however, speak of generalizability only if one has indicated to what universe the generalization will be made. According to Cronbach et al (1963:145),

No assumptions should be made about the content of the total universe nor the statistical properties of the test scores within it, but specific requirements to be met are:

1. The universe must be described unambiguously, so that it is clear what conditions fall within the universe.

2. Conditions are experimentally independent: the person's score i does not depend on the fat that he has, or has not been previously observed under what conditions.

3. Scores X_{pi} are numbers on an interval scale.

The candidate's universe score is defined as the mean of his sampling distribution of the means over all sample tests in the universe (Webb et al, 1983). To ask if a candidate's obtained score is reliable, is tantamount to asking how confident one can be in generalizing from the obtained score in hand to some defined class of observation to which the sample behaviour belongs (Cronbach et al, 1963). Considering test scores in this way is a departure from the classical test theory where it is thought that each test has a true score and belongs to some family of parallel tests and any test from this "family" of parallel tests must be equivalent in statistical characteristics (Lord and Novick, 1968). Rajaratnam et al (1965) argue that application of a random sample model in test construction produces tests with items which would represent the domain without assuming equivalence as is the case in the classical test theory. However, it has been argued that this model is inadequate because it does not guarantee representation of each type of behaviour in the domain. It has been suggested that for a sample to adequately represent the universe, it must duplicate or reproduce the essential characteristics of the universe in the proper proportions (Cronbach, 1970). To be able to determine the extent of generalizability from a test whose items have been randomly sampled from the domain, computation of coefficient ($\hat{\alpha}$) as an index of generalizability was suggested by Rajaratnam et al (1965).

Because of the shortcomings of the random sample model, the stratified random sample model of test construction had been suggested by the same authors. The latter calls for stratification of behaviour within a domain, each stratum having items which would eventually appear on a test. The test constructor then systematically randomly selects items from each stratum after deciding on the number of items needed from each stratum. Computation of stratified alphas ($\hat{\alpha}$s) for this model is the index of generalizability which is supposed to provide a better estimate of

2. Conditions are experimentally independent: the person's score i does not depend on the fat that he has, or has not been previously observed under what conditions.

3. Scores X_{pi} are numbers on an interval scale.

The candidate's universe score is defined as the mean of his sampling distribution of the means over all sample tests in the universe (Webb et al, 1983). To ask if a candidate's obtained score is reliable, is tantamount to asking how confident one can be in generalizing from the obtained score in hand to some defined class of observation to which the sample behaviour belongs (Cronbach et al, 1963). Considering test scores in this way is a departure from the classical test theory where it is thought that each test has a true score and belongs to some family of parallel tests and any test from this "family" of parallel tests must be equivalent in statistical characteristics (Lord and Novick, 1968). Rajaratnam et al (1965) argue that application of a random sample model in test construction produces tests with items which would represent the domain without assuming equivalence as is the case in the classical test theory. However, it has been argued that this model is inadequate because it does not guarantee representation of each type of behaviour in the domain. It has been suggested that for a sample to adequately represent the universe, it must duplicate or reproduce the essential characteristics of the universe in the proper proportions (Cronbach, 1970). To be able to determine the extent of generalizability from a test whose items have been randomly sampled from the domain, computation of coefficient ($\hat{\alpha}$) as an index of generalizability was suggested by Rajaratnam et al (1965).

Because of the shortcomings of the random sample model, the stratified random sample model of test construction had been suggested by the same authors. The latter calls for stratification of behaviour within a domain, each stratum having items which would eventually appear on a test. The test constructor then systematically randomly selects items from each stratum after deciding on the number of items needed from each stratum. Computation of stratified alphas ($\hat{\alpha}s$) for this model is the index of generalizability which is supposed to provide a better estimate of

generalizability, thus implying that stratified sampling is more generalizable to the total universe of interest.

The theory of generalizability is based on the fact that scores on tests constructed from an explicitly defined universe of item content, give an unbiased estimate score of the total universe (Cronbach et al, 1963). The problem test designers come up against is the construction of comprehensive criterion referenced tests whose scores are directly interpretable in terms of specified performance standards, and be generalizable to the whole domain implied by instructional objectives. Generalization is more crucial in this kind of testing than in norm referenced testing because in the case of the former, one is more concerned about making absolute interpretations in order to assess mastery. Absolute interpretations refer to interpretations of test scores made without reference to the scores of other examinees who took the test (Cronbach, 1971). In cases where the relative order of individuals is of primary concern (i.e. in norm referenced tests) relative interpretations are important (Cronbach 1971). In these cases the value of the examinees' scores usually carries less absolute meaning and must be interpreted in terms of the norm group. While generalizability may be important in norm referenced tests, item sampling which maintains relative order from sample to sample is of more concern.

Absolute interpretation of scores is especially important in criterion referenced tests because in norm referenced tests as long as the selection of items is carefully made, test scores will discriminate among examinees.

Bloom (1968) defined aptitude as the ability to learn tasks and individuals differ in their aptitudes to learn these tasks. Carroll (1963) suggested that if achievement measures are both reliable and valid, the correlations between aptitude and achievement can be up to +.70 or better. Bloom (1968) stated that if students who vary in their aptitude went through an instructional programme of the same quality, but the instructional procedures meet the needs of each individual and allow individuals to progress at their own pace, the majority of these individuals could be expected to achieve mastery of the subject matter. The correlation

between aptitude and achievement in this kind of set-up eventually reduces to zero. Glaser (1968:172) remarked that:

> individual instruction requires the fine honing of instructional procedures so that a student seeks and achieves mastery by proceeding along a path to a large extent dictated by the individual student's requirement.

Caroll (1963), defining aptitude in relation to mastery, suggested that aptitude could be regarded as the amount of time spent by the learner to acquire mastery. If this be the case, and given enough time all students should be able to attain mastery because they would learn tasks at their own pace, and the majority of them would eventually attain mastery of each learning task, albeit others achieving mastery sooner that others.

In a programme of mastery learning, one may wish to assess pupils' progress, growth, development, or change. The process of measuring the student's terminal performance during and at the end of instruction has been called performance assessment by Glaser (1962).

As already stated, two types of tests used to measure mastery of instructional objectives are criterion referenced tests which are deliberately constructed to yield scores that are directly interpretable in terms of specified performance standard and norm referenced tests which use relative standards.

Criterion referenced tests used as achievement or mastery tests must be generalizable to the task domain as specified by the instructional objectives (Glaser and Nitko, 1971). It is also important that they should adequately sample the domain of subject matter about which inferences are to be made (Shavelson et al, 1981).

The major problem is constructing criterion referenced tests has been to select items that adequately sample behaviour defined by the domain at hand, while at the same time keeping the relative characteristics of the domain to which the test performance will be generalized (Glaser and Nitko, 1971).

Cronbach (1970), discussing content validity suggested that a good test is not guaranteed merely by assembling "good items". It is the ensemble of items that must be considered, and if one is to judge whether the ensemble samples the right kinds of behaviour there must be clear specifications of the behaviours. If this can be done carefully then the test should have content validity. Content validity, in this case, is evidenced by showing how well the content of the test samples the subject matter about which conclusions are to be drawn. Reliability thus becomes a question of *accuracy in generalization.*

Generalizability to the universe of content is of paramount importance in any instructional programme. Carver (1970) observed that very seldom in measuring achievement are teachers exclusively interested in performance on just those items that are actually administered on a test. They invariably have some larger universe of content in mind. The implicit objective is to generalize over the entire content domain.

If we indeed assume that items are samples from the domain of relevant tasks, then the problem of generalizing an individual's performance to the task domain can be thought of as an item sampling problem. It is in this context that two methods for sampling test items have been suggested:

1. simple random sampling of items from a well defined domain; and

2. representative or stratified random sampling from the same domain.

A simple random sample model requires that items appearing on a test be randomly sampled from the universe which has been explicitly defined. According to Rajaratnam et al (1965) such a test need not necessarily have equivalent items.

Cronbach et al (1963) suggested two kinds of alphas as indices of generalizability that could serve as unbiased estimates of the generalizability to the domain. The first proposed alpha ($\hat{\alpha}$) was to be used when a test had been assembled by means of simple random sampling from a defined universe. The other alphas ($\hat{\alpha}_s$) when stratified random sampling had been used.

Objections had been raised to using the simple random sample model in test construction. Cronbach et al (1963:159) for instance, stated that some objections raised were:

(a) that the universes to which one might refer are usually vaguely defined and not denumberable,

(b) that strict random sampling from a pool of items or judges rarely occurs.

However, employing Hively et al's (1968) item form analysis of the universe would seem to answer these two objections. As a further reply to such criticism, Cronbach et al (1963:160) argues that

> The absence of true random sampling from a pool of items or judges is unfortunate but no more so than in the ubiquitous studies that make statistical inferences from persons who are not chosen in a strictly random fashion.

The simple random sample has been found to be unrealistic in that when building a test, items usually come from a domain which is stratified in some way. To ensure content validity one would like each stratum in the domain to be proportionally represented. Rajaratnam et al (1965) suggested this as another "model in which a test is considered to have been formed by stratified sampling of items". The coefficient ($\hat{\alpha}_s$) already mentioned above is an index of generalizability computed on a stratified random test to estimate the universe score, $E(\sigma^2_{Mt})$. The same authors, however listed certain assumptions to be met before this could be done (Rajaratnam, 1966:43).

1. that there is a universe of items, divided into fixed strata h(h = 1, . . . ,m);

2. that such stratum contains an infinite number of items;

3. that there is a sampling plan which specifies the number of items k to be drawn from a particular stratum h;

4. that there is an indefinitely large potential family of stratified parallel tests that could be constructed by randomly sampling within strata in conformity with the sampling plan, the test in hand being a member of that family.

The heart of the generalizability theory is the assumption that a person's score on a test can be used to infer the person's universe score.

To determine the unbiased estimate of the coefficient of generalizability $E(\sigma^2_{Mt})$ when a test is not stratified, the following computation was suggested by Rajaratnam et al (1965):

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\Sigma S_i^2}{S^2_t}\right)$$

or where an item is scored 1 or 0,

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{n \Sigma P_i(1-P_i)}{(n-1) S^2_t}\right)$$

P_i is the proportion of examinees answering item i correctly. S_i^2 is the item variance.

The formula suggested by the same authors to estimate $E(\sigma^2_{Mt})$ when a test is stratified is:

$$\hat{\alpha}_s = 1 - \sum_h \frac{1}{k_h - 1} \frac{(k_h S^2_{ih} - S^2_h)}{S^2_t}$$

Where:

h = strata in the test

k_h = number of items to be drawn from a stratum h

i = item on a test

S_t^2 = item variance

S_t^2 = total test variance

S_{ih}^2 = variance of an item within a stratum

and when items are scored 1 or 0, the sample variance may be calculated as

$$\frac{n}{n-1} P_i (1-P_i)$$

where n is the number of items in the test.

Stratified alpha ($\hat{\alpha}_s$), like unstratified alpha ($\hat{\alpha}$), is also an approximation to $E(\sigma^2_{Mt})$. Both $\hat{\alpha}$ and $\hat{\alpha}_s$ have the same interpretation. Both are intraclass correlations and are the means of all the possible split half coefficients. The split halves for $\hat{\alpha}$ are obtained by dividing a test in half in all possible ways. The same interpretation is true for $\hat{\alpha}_s$ only that it is the mean of split half (intraclass) coefficients for the test if the splitting process takes into account the strata of the test (Cardinet et al, 1981). To determine which of the two, $\hat{\alpha}$ or $\hat{\alpha}_s$ has a better index of generalizability, it is necessary to compare their "signal to noise" (S/N) ratios, computed by $\hat{\alpha}/(1 - \hat{\alpha})$ and $\hat{\alpha}_s/(1 - \hat{\alpha}_s)$ respectively. The one with a higher ratio has a better index of generalizability.

As to whether stratification should be based on the homogeneity of item content or indices of item difficulty (that is, putting items of the same difficulty index in one stratum), Cronbach et al (1965:311) stated that:

> Stratifying on content is clearly more important than stratification on difficulty, both in construction and test analysis. The so-called difficulty factors that have received so much attention from some test theorists prove to have very little influence on coefficients unless r_w (inter-item tetrachoric correlation) is unrealistically high.

Although the theory of generalizability appeals to test constructors, very few attempts have been made to put it into practice because of its mathematical and technical applications, and also the difficulty inherent in satisfying the assumption that the universe to which generalization would be made must be explicitly defined.

Purpose of the Study

This study investigated the magnitude of generalizability of test scores to the universe when applying different models for constructing achievement tests which would give estimates of examinees' universe scores. In particular, the study investigated empirically whether there was a difference in coefficient of generalizability when tests were constructed by simple random sample model and when stratification was taken into account, and to determine if stratified or a simple random sample model should be used when constructing criterion referenced tests designed to assess mastery of instructional objectives. The study looked into the equivalence of the two tests and ascertained the feasibility of applying generalizability theory to test construction.

Research Hypotheses

This study was carried out to answer the following questions:

1. Which model of test construction (simple random or taking stratification into account) would produce a better coefficient of generalizability as measured by the test's coefficient of generalizability index to enable a more accurate interpretation of raw scores in relation to the universe score.

2. Would the two types of tests, simple random and stratified, constructed from the same item pool (item forms) be equivalent? That is, would they have equal means and variances and high intercorrelation to meet the classical test theory definition of equivalence?

3. Is the application of generalizability theory to achievement test construction feasible?

Procedures

The instructional objectives in mathematics selected for the study were taken from the primary school syllabus whose domains of subject matter were explicitly defined by item forms. The mathematics curriculum for grades one through to seven from which the instructional objectives for the study were chosen is comprised of approximately 400 instructional objectives which have been divided into 80 units for instructional purposes. Specific item forms for this study included application items and were grouped into item form units which shared a common content in multiplication, fractions, division, difficulty levels, etc.

An example of the hierarchical instructional objectives used for the study and their item generation rules which explicitly defined the subject matter is depicted in Table 1.

Table 1
Hierarchical Instructional Objectives for Level F Primary Mathematics in Grade 7

1.	Given a two-digit number times a two-digit number, the student multiplies using the standard algorithm.
2.	Given a three-digit number times a two-digit number, the student multiplies using the standard algorithm.
3.	Given a whole number and a mixed decimal to hundredth's as factors the student multiplies. LIMIT: whole number part < 100
4.	Given product of two pure decimals, $< .99$, the student shows the equivalent in fractional form and converts product to decimal notation, compares answers for check.
5.	Given a multiple step word problem requiring multiplication skills mastered to this point, the student solves (< 3 steps).

Two achievement tests - one constructed by a simple random procedure and the other by stratified random sampling procedure - from the same instructional objectives were assembled from test items generated from their item forms after the subject matter had been precisely defined by their item forms (Hively et al, 1968). The design is summarized in Table 2. Experimental subjects for the study were mathematics students in their final year of primary school. The thirty students used for the study were not randomly selected but the teacher selected a group which was mixed in cognitive ability and included both boys and girls.

The mathematics class was divided into two groups of fifteen students each. The procedure for forming groups was to take every other student (as they sat in class) and call them Groups I and II, respectively. In order to confound the order effects, on the first day of administration, the first group got the stratified form, and the second group got the simple random form of the test.

On the second administration, which was one week after the first, the tests were administered in reverse order. Group I took the random form and Group II took the stratified form. Matched data (depicted in Appendix I) of the study was thus collected on both types of tests. Both forms had 32 items and every student answered all the questions.

Table 2
Test Form

	Stratified	Random
Group I	First day of Administration N = 15	Second day of Administration N = 15
Group II	First day of Administration N = 15	Second day of Administration N = 15

Results

The analysis of the two tests indicated that the distribution of test scores were both negatively skewed. The simple random test was more negatively skewed than the stratified test as can be seen in Figure 1. The frequency polygons for the two tests are reported in Figure 1 as well. Descriptive statistics for the two tests are shown in Table 3. There was a slight tendency for the mean of the stratified test to be higher than the mean of the random test. This difference was not statistically significant at the 5 percent level ($t = -1.414, df = 29$).

Generally, over the whole test, the mean item difficulty is slightly higher for the stratified test. Appendix II shows the distribution of item variances, mean item difficulty and discrimination indices within the ten strata used for the study. The table also shows that the main difference in item variability occurred on items from stratum 10. The simple random test had more variability.

As can be seen from Table 3, the total test score variance was larger for the simple random test. The difference between the two variances was tested by t-test for related samples (Glass and Stanley, 1970). This difference was significant at the .05 level ($t = 5.29$, df. $= 28$). Examination of the two distributions also indicates that, other than the one extreme score on the simple random test, the examinees found both tests easy, as could be seen from the distributions in Appendix I. One could also interpret this as being mastery of the instructional objectives which were used for this study.

Figure 1
Frequency of polygons of the two tests

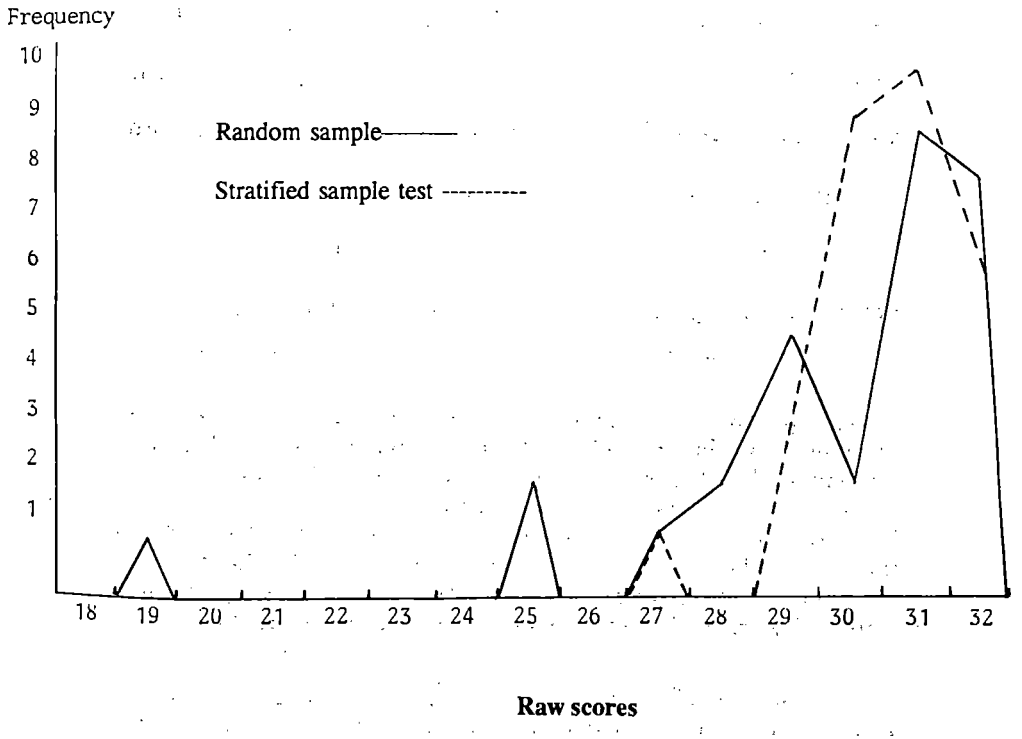

Table 3
Results Of Test Statistics Unbiased Estimates

Statistics	Random Tes	Stratified Test
No, of subjects	30	30
Mean	29.73	30.50
Total Test Variance	7.941	1.362
Standard Deviation	2.818	1.167
Mean Item Variance	0.063	0.040
Mean Item Difficulty	0.930	0.950
Intercorrelation	0.154	
Maximum Correlation	0.923	

Because the two distributions were negatively skewed Carroll's (1961) maximum correlation coefficient was computed which indicates the maximum value that the correlation coefficient can reach given the two distributions. This maximum value was 0.923. The observed inter-correlation (Pearson-product moment) between the two tests was 0.154. The value of the correlation plus the fact that the variances were different indicates that the two tests were not equivalent, at least in the statistical sense implied by the classical test measurement model.

Coefficient alpha ($\hat{\alpha}$) and alphas ($\hat{\alpha}_s$) as indices of generalizability were computed on both tests and are presented in Table 4. The alphas on the two tests were converted to their signal/noise ratio as suggested by

Cronbach et al, (1965) so that their coefficients of generalizability could be compared. The conversion used the formulas S/N ratio = $\hat{\alpha}/(1-\hat{\alpha})$, for the random test and $\hat{\alpha}_s/1-\hat{\alpha}_s$ for the stratified test when stratification was taken into account on both tests. The results of this conversion are shown in Table 4 where the signal to noise ($\hat{\alpha}_s$) ratio for the random test is higher than ($\hat{\alpha}_s$) for the stratified test, but within test analysis the signal to noise ratios for ($\hat{\alpha}_s$) are higher than those for ($\hat{\alpha}$) when stratification is taken into account.

Table 4
Results of Coefficients of Generalizability

	Random Tes	Stratified Test
Coefficient Alpha ($\hat{\alpha}$)	0.764	0.012
Coefficient Alphas ($\hat{\alpha}_s$)	0.872	0.079
Signal/Noise Ratio ($\hat{\alpha}$)	3.237	0.012
Standard Deviation ($\hat{\alpha}_s$)	6.813	0.086

Between the two tests the random test had higher coefficients of generalizability to the domain of items. The contributory factor to this was the fact that the random test had a higher variance of 7.941 as compared with 1.362 for the stratified test as depicted in Table 3. Assuming that stratification were to be done on the random test and ignoring stratification on the other test so that both models could be applied for each test, computation of $\hat{\alpha}$ and $\hat{\alpha}_s$ for each test (i.e. computing random and stratified alphas for both tests), as indicated in table 4 showed that the coefficient of generalizability increased in both cases when *stratification* was taken *into account*. To determine how much longer a test would need to be increased to reach the coefficient of generalizability

when stratification was taken into consideration, the S/N ratios were used where the value of signal/noise ratio ($\hat{\alpha}_s$) for each test was divided by the value of the signal noise ratio ($\hat{\alpha}$) and the result multiplied by the original number of test items which was 32 for each test. For the random test of 32 items, to increase the coefficient of generalizability from 0.764 to 0.872 the test would need to be increased to 68 items. For the stratified test to increase in coefficient of generalizability from 0.012 to 0.079, the test would need to be increased to 230 items. The results then demonstrated that one would get a better index of generalizability if tests took into account stratification of item forms when constructing criterion referenced achievement tests.

Summary And Conclusions

The study investigated empirically:

(1) whether there was a difference in coefficient of generalizability when achievement tests were constructed by simple random sample or stratified random sample model;

(2) if the two tests met the classical test theory criteria for equivalence;

(3) the applicability of generalizability theory to the construction of criterion referenced achievement tests.

Two achievement tests were constructed each having 32 items. One test was constructed by simple random sampling from a defined domain and the other by stratified random sampling. For these two mastery tests the results indicated that the random test compared with the stratified test had a higher coefficient of generalizability, although the within comparison analysis indicated that taking strata into account increased the generalizability coefficient of both tests, which means that if test constructors stratified test items from a randomly assembled pool of items, the generalizability of test to the universe could be enhanced.

The results on the whole indicated that most of the students had mastered the instructional objectives by the time they responded to the items on the two tests.

The results also indicated that the two tests were not equivalent, judging by their test variances and their intercorrelation coefficient, confirming that if two types of tests - random and stratified - are constructed from the same item forms they would not necessarily meet the classical test theory criteria for test equivalence. The study also demonstrated that the application of generalizability theory is feasible when constructing criterion referenced achievement tests.

REFERENCE

Bloom, B.S. (1968). Learning of Mastery. U.C.L.A. Evaluation Comment, 1:2.

Cardinet, J, Tourneu, Y. & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement*, 18, 183-204.

Carroll, J. (1961). The nature of data, or how to choose a correlation coefficient. *Psychometrika,* 26, 347-372.

Carroll, J. (1963). A model of school learning. *Teachers College Record,* 64, 723-733.

Carver, R. (1970). Special problems in measuring change with psychometric devices. *Evaluative Research,* A.I.R. 48-63.

Cronbach, L.J. (1970). *Essentials of Psychological Testing,* (3rd Edition), New York: Harper and Row.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (ed.). *Educational Measurement,* Washington, DC: American Council on Education, 2nd Edition, 443-507.

Cronbach, L.J. Rajaratnam, N. & Gleser, G.C. (1963). Theory of generalizability; a liberalization of reliability theory. *British Journal of Statistical Psychology,* 16, 137-163.

Cronbach, L.J. Rajatnam, N. & Gleser, G.C. (1965). Alpha coefficients of stratified-parallel tests. *Educational and Psychological Measurement,* 25, 291-311.

Glaser, R. (1962). Psychology and instructional technology. *Training Research and Education.* Pittsburgh, 1-30.

Glaser, R (1968). The design and programming of instruction - the schools and the challenge of innovation. Supplementary paper No. 28, New York: Committee for Economic Development, 166-215.

Glaser, R & Nitko, A.J. (1971). Measurement in learning and instruction. In R.L. Thorndike (ed.). *Educational Measurement,* Washington, D.C.: American Council on Education, 2nd Edition, 625-670.

Glass, G.V. & Stanley, J. (1970) *Statistical methods in education and psychology,* Englewood Cliffs: Prentice-Hall.

Hively, W. (1966). Preparation of a programmed course in algebra for secondary school teachers. A report to the National Science Foundation under NSFG 25164, Amendments 2, 3, and 4.

Hively, W., Patterson, H & Page S. (1968). An "Universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement,* 5, 275-290.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Addison Wesley, Massachusetts.

Mager, R.F. (1962). *Preparing instructional objectives.* Pearson Publications, Belmont, California.

Rajaratnam, N., Cronbach, L.J. & Gleser, G.C. (1965). Generalizability of stratified-parallel tests. *Psychometrika*, 30, 39-66.

Shavelson, R.J. & Webb, N.M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology,* 34, 133-166.

Shavelson, R.J. & Webb, N.M. (1989). Generalizability theory. *American Psychologist,* 44, 922-932,

Webb, N.M., Shavelson, R.J., & Maddahian, E. (1983). Multivariate generalizability theory. In L.J. Fyans, Jr. (ed.). *Generalizability theory: inferences and practical applications.* San Francisco: Jossey-Bass, 67-81.

APPENDIX I

Table 5
Test Raw Scores with a maximum of 32

Students	Random	Stratified
1	29	32
2	32	30
3	25	31
4	30	32
5	32	32
6	31	29
7	32	31
8	27	30

9	32	31
10	32	31
11	29	27
12	31	32
13	31	30
14	31	30
15	25	31
16	29	31
17	31	31
18	32	29
19	32	30
20	29	32
21	30	31
22	31	30
23	32	30
24	29	31
25	31	31
26	30	32
27	31	30
28	28	30
29	28	29
30	19	29

Student No	Random	Stratified
1-15	2nd Admin. $\overline{X} = 29.93$	1st Admin. $\overline{X} = 30.60$
16-30	1st Admin. $\overline{X} = 20.47$	2nd Admin $\overline{X} = 30.40$

APPENDIX II

Item Means, Variances, and Discrimination Indices

Item No.	Stratum 1			Stratum 2			Stratum 3			Stratum 4			Stratum 5			Stratum 6			Stratum 7			
	12	23	Sub. Tot.	2	25	29	Sub. Tot.	28	31	Sub Tot.	8	18	Sub Tot.	4	17	Sub. Tot.	0	0	Sub. Tot.	7	19	22	Sub. Tot.
Random																							
X	.97	.97	1.9	.77	.93	.93	2.63	.93	.93	.186	.93	.90	1.83	.97	1.00	1.97	—	—	—	.93	.97	.97	2.87
s^2	.03	.03	.06	.183	.067	.067	.317	.067	.067	.134	.067	.093	.160	.03	.00	.03	—	—	—	.067	.030	.030	1.27
Dis.Index	.43	.00	—	.72	.58	.43	—	.43	.43	—	.58	.58	—	.43	.00	—	—	—	—	.43	.00	.43	—
Item No.	2	5	Sub. Tot.	4	23	Sub. Tot.	20	32	Sub Tot.	18	27	Sub Tot.	7	16	Sub. Tot.	3	11	Sub. Tot.	17	21	Sub. Tot.	
Stratified																						
X	1.00	1.00	.2	1.00	.93	1.93	.83	.93	1.76	.97	.97	1.94	.93	1.00	1.97	.93	.97	1.9	1.00	1.00	2	
s^2	.00	.00	.00	.00	.067	.067	1.46	.067	.213	.03	.03	.060	.067	.00	.067	.067	.03	.097	.00	.00	0	
Dis.Index	.00	.00	—	.00	.00	—	.72	.43	—	.00	.00	—	.00	.00	—	.43	.43	—	.00	.00	—	

Item No.	Stratum 8				Sub Tot.	Stratum 9		Sub. Tot.	Stratum 10														Sub. Total
	3	15	20	30		5	16		1	6	9	10	11	13	14	21	24	26	27	32	
Random																							
X	.90	.97	.93	.97	3.77	1.0	.97	1.97	.87	.90	.87	.87	.90	.97	.90	.93	.97	.97	.93	.93	10.8	
s^2	.093	.030	.067	.030	.220	0	.03	.03	.117	.093	.117	.117	.093	.03	.093	.067	.117	.03	.067	.067	1.008	
Dis.Index	.58	.43	.58	.43	—	.00	.00	—	.58	.67	.72	.87	.58	.43	.43	.58	.58	.43	.43	.58	—	

Item No.	Stratum 8		Sub. Tot.	Stratum 9		Sub. Tot.	Stratum 10													Sub. Total
	15	19		1	26		6	8	9	10	12	13	14	22	24	25	28	29	30	31	
Stratified																					
X	1.00	.90	1.90	.97	1	1.97	1.00	.90	.93	.80	.93	.97	.90	.93	.97	.90	.90	1.00	.97	.97	13.17
s^2	.00	.093		.03	.00	.03	0.00	.09	.067	.093	.067	.03	.093	.067	.030	.093	.093	.00	.03	.03	.786
Dis.Index	.00	.43	—	.00	.00	—	0.00	.58	.43	.00	.43	.43	.58	.00	.00	.43	.43	.00	.43	.43	—