The Use of Generalizability Theory for Assessing Relations among Items within Domains in Diagnostic Testing

George B. Macready
University of Maryland

The purpose of this study was to describe a procedure based on generalizability theory for assessing the adequacy of item groupings generated by means of a domain-referenced testing system. The procedure is based on an exploration of the extent to which items are grouped into homogeneous sets with similar levels of difficulty, when the sets are defined on the basis of a logical analysis of the perceived skills underlying the items. In addition, an example of the use of the developed analytic procedure is presented.

Educational achievement tests are generally constructed to be representative of "important" content areas within the subject area of interest. However, in the construction of such tests, the subject matter is often not specified in detail. This can result in confusion regarding the manner in which test items are selected for a test or the kinds of statements that can be appropriately made about examinees based on their test performance.

One approach to creating educational achievement tests that have clear content structure is domain-referenced testing (Hively, 1974; Hively, Patterson, & Page, 1968). Through this approach, subsets of items called *domains* are defined in terms of operationally stated rules called *item form rules,* which allow for an explicit description of the complete set of items that could be written (the item "universe"). Hively (1970), in discussing the nature of domain-referenced testing, stated:

> A domain-referenced testing system consists of rules for sampling items from a domain and administering them to an individual . . . in order to obtain estimates of the probability that the individual . . . could answer an arbitrarily chosen item from the domain . . . Precise definition of a domain and its subsets provides for an exact diagnosis of performance. Thus . . . once we have diagnosed a student with respect to a defined domain, we may be able to predict his behavior in natural situations which have properties in common with test situations within the domain. (p. 3)

The purpose of this paper is to present a procedure based on generalizability theory for assessing the effectiveness with which domain-referenced testing systems provide acceptable groupings of items. Also, the procedure is proposed as useful for identifying which, if any, of the generated domains may be in need of modification.

With the adoption of a domain-referenced approach to testing, it is possible to assess relations between an examinee's response to a set of randomly selected items from a specified domain and the proportion of all items in that domain to which the examinee could (theoretically) correctly respond. Within the area of diagnostic testing in which assessments are made regarding the presence of specific skills or traits, it may be desirable to at-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 2, Spring 1983, pp. 149–157

0146-6216/83/020149-09$1.45

tempt to define item domains narrow in scope. This makes it more reasonable to assume that the necessary and sufficient skills required to answer any item in the domain is the same (or similar) for all items. Thus, except for error of measurement, examinees may be expected either to be able to respond correctly to all items in a domain or to be incapable of responding correctly to any of the items in that domain. When domains are effectively defined in terms of scope, two properties may frequently be expected to be present. First, all items within a domain will have comparable levels of item difficulty and, second, a high level of homogeneity among items will exist (Loevinger, 1947).

If these conditions are approximated, estimates of examinees' expected item scores across all items in a domain (here called "domain scores") will tend to be accurately estimated from their observed scores on even a small number of randomly selected items from the domain. In addition, estimated domain scores should effectively identify which examinees have acquired the necessary and sufficient skills to respond correctly to all items within the domain. More extensive discussions of desired interitem relations within domains which are appropriate for diagnostic testing are provided by Harris (1974) and Macready and Merwin (1973).

If a domain shows what is considered to be too little homogeneity among items or too much variability in item difficulty, it may be desirable to try to redefine the item from rules defining the domain such that the scope of item content for the domain is narrowed. In other cases, it may be desirable to broaden the scope of the domain. This is especially true if two or more domains of items may be collapsed or restructured without greatly reducing the extent to which the desired criteria are met. Such an approach would appear to be preferable to assessing the adequacy of specific items from a domain, since specific modification or deletion would result in ambiguity regarding the resulting item universe (see Kane, 1982; Lumsden, 1976).

Attempts to redefine domains might incorporate a conceptual assessment of item homogeneity using procedures suggested by Hartke (1978) to identify possibly viable domain revisions. Once a domain structure has been established (whether original or revised), its adequacy can be addressed through an assessment of item homogeneity and variability in item difficulty using the procedures presented in this paper.

To determine the accuracy with which examinees' domain scores are in fact estimated with item samples of a specified size, a "generalizability coefficient" may be used. The coefficient is an estimate of the ratio of domain score variance to expected observed score variance and provides an index of the relative accuracy with which scores obtained on a randomly sampled subset of items of a given size permit generalization to the universe of elements from which the subset was sampled. Cronbach, Gleser, Nanda, and Rajaratnam (1972) suggest that the undifferentiated error component in classical reliability theory might more reasonably be differentiated. Such differentiation may provide a better understanding of sources of error and their possible effects. Under this approach the total variance among item scores is attributed to multiple sources by the incorporation of a multifactor analysis of variance model. Cronbach et al. (1972) have called such an analysis of the components of variance a "generalizability study." The application of such studies in the assessment of criterion-referenced tests has been given considerable attention by Brennan (1978, 1979), Brennan and Kane (1977a, 1977b), and Cardinet, Tourneur, and Allal (1976).

Method

The domain-referenced testing system used to exemplify the procedure presented in this study was the section of Honeywell's "Arithmetic Test Generation Program" (ATG) dealing with multiplication of whole numbers (Patterson & Vierling, 1970). The ATG is a revision of the testing system described and used by Hively et al. (1968).

In the area of multiplication of whole numbers, the ATG has grouped items into 20 separate domains. Items from six of the domains were used in this study. In the selection of the six domains that were considered, an attempt was made to obtain domains whose average item difficulties were neither extremely high nor low. This was done to

prevent reduction in variability due to extreme difficulties from limiting the effectiveness of the generalizability study. The general item characteristics found within each of the six domains are listed in Table 1.

Ten items were randomly generated from each domain with the restriction that half the items had four-digit multiplicands and half had three-digit multiplicands. The 60 items generated constituted the specific content of the test. Students were tested on each of two consecutive days. (This was necessary due to practical considerations and not because "testing occasions" was of particular interest as a facet in the analysis). Items from three domains were administered on the first day, and the items from the remaining domains on the subsequent day. The sets of items from the various domains were administered on the two days in a counterbalanced manner in order to eliminate testing occasion. A comparable number of the students from each class were randomly assigned to each of the six forms of the test defined by the counterbalancing procedure.

During each day of testing, one randomly chosen item from each domain being considered was presented twice to the students. The specific items that were presented twice differed across forms of the test. (However, the six test forms contained the same items). The use of different items for two presentations on each test form was done to obtain samples of "repeated" items from each domain which were more representative of their respective domains. The specific location on the test of the repeated items was held constant from form to form and from testing period to testing period.

Subjects were 260 students, equal numbers of whom were randomly selected from each of 10 fifth-grade classrooms in the Minneapolis Public School System. The actual sampling of classrooms was done by school so that when a particular school was selected, all of the fifth-grade classrooms in that school were involved in the study.

Table 1
Characteristics Describing Items Found in the Domains Studied

Domain No.	Prototype Item	Item Format	Item Characteristics
10	824 ×21	A ×B	2 digit multiplier; no carry
12	432 ×40	A ×B	2 digit multiplier; multiple of 10
13	347 ×23	A ×B	2 digit multiplier; easy carry
15	8647 ×69	A ×B	2 digit multiplier; hard carry
17	627 ×204	A ×B	3 digit multiplier with middle digit equal to 0
18	472 ×361	A ×B	3 digit multiplier with no digits equal to 0

Results

Assessment of Generalizability

A generalizability study was conducted to determine the variance associated with each of the following facets:

c—classrooms at the fifth-grade level (10 classrooms were selected);
d—domains of multiplication items (six different domains of multiplication items were included);
n—number of digits found in the multiplicand of each item (multiplicands with three and four digits were included);
i:dn—items nested within the domains with a given number of digits in their multiplicands (five items were randomly sampled from each of 12 *dn* categories); and
s:c—students nested within classrooms (26 students from each class were randomly sampled).

The estimated variance components for each source of variation, $\hat{\sigma}^2(\ )$, resulting from the above design, along with their corresponding estimated standard errors (see Searle, 1971, p. 417) and the proportions of total variance accounted for, $\hat{\sigma}^2(\ )/\hat{\sigma}^2_{\text{total}}$, are presented in Table 2.

In evaluating the values of the estimated variance components obtained in this study, it should be pointed out that all of their estimated standard errors are small in magnitude; thus, it is reasonable to assume that the estimated variance components are reasonably stable (see Searle, 1971, p. 417). Note that the single source of variation with the largest variance is *s:c* × *i:dn,e*, which accounts for 47% of the total variance. This result would seem to indicate that one or more domains contain items which do not result in highly homogeneous response outcomes for students, thus deviating from the desired property of item homogeneity within each domain. In turn, this would seem to indicate the need for separate assessments of each domain with respect to item homogeneity. The remaining four sources of variation, which account for one or more percent of the total variance, are *c*; *s:c*; *d*; and *s:c* × *d*, which in combination account for 51% of the total variance. The first three of these relatively large components of variance indicate that there is considerable variability among means related to the various levels of classrooms, students (within classrooms), and domains. The large variance obtained for the last of these sources suggests that there are differences in the relative difficulty of items from various domains occurring for various students. In the case of all four of the above sources, the results obtained are not necessarily required for an effectively structured domain-referenced testing system, although they are compatible with such a system.

Further insights into the underlying structure of the domain-referenced testing system were provided by looking at those remaining sources of variation that accounted for only a small proportion of the item score variance. For example, the small amount of variation related to "number of digits" in the multiplicand (i.e., either three or four digits) provides support for the contention that this variable does not need to be restricted to a single level in order to attain comparability of item difficulties within domains. This conclusion is given further support by the fact that none of the interaction effects involving number of digits accounted for sizable portions of the total item score variance. These findings suggest that the specification of number of digits in the multiplicand within the item form rules for the domains is not necessary. One possible implication of this assessment might be that items with three-digit multiplicands may effectively be used alone to estimate students' domain scores. Such a practice could prove quite valuable because of the savings in time provided by students working the shorter items.

Another source of variation which accounted for a very small portion of the total variance was *i:dn*. This suggests that the differences in item difficulties among items within a given item set (i.e., items falling within the same domain and with the same number of digits in their multiplicand) tend to be small. This conclusion is supported further by separate nested analyses that were run on each set of items.

Table 3 presents both the absolute $[\hat{\sigma}^2(i)]$ and the relative $[\hat{\sigma}^2(i)/\hat{\sigma}^2_{\text{total}}]$ variances attributed to item difficulties within each set of items. (These statis-

Table 2
Generalizability Study Analysis of Variance

Sources of Variation*	df	MS	$\hat{\sigma}^2(\ )$	Estimated Std. Error of $\hat{\sigma}^2(\ )$	$\hat{\sigma}^2(\ )/\hat{\sigma}^2_{total}$
c	9	39.75	.0226	.0109	.088
s:c	250	4.49	.0728	.0067	.284
d	5	41.40	.0156		.061
dc	45	0.45	.0005	.0004	.002
s:c × d	1250	0.31	.0189	.0012	.074
n	1	9.80	.0012		.005
cn	9	0.18	.0000	.0001	.000
s:c × n	250	0.15	.0009	.0004	.004
dn	5	1.06	.0005		.002
cdn	45	0.12	-.0001[a]	.0002	.000
s:c × dn	1250	0.12	-.0007[a]	.0010	.000
i:dn	48	0.49	.0014	.0004	.007
i:dn × c	432	0.14	.0007	.0004	.003
s:c × i:dn,e	12000	0.12	.1210	.0015	.472

*Where: c = Classrooms (a random variable)[b],
s = Students (a random variable)[b],
d = Item Domains (a fixed variable),
n = Number of digits in the multiplicand (a fixed variable),
i = Items (a random variable),
e = Residual error and
$\hat{\sigma}^2_{total} = \Sigma\hat{\sigma}^2(\ )$ for the s:c × i:dn design analysis.

[a] This variance component is assumed to be equal to zero

[b] Although the levels of this variable were not randomly selected, for the purpose of this study the variable is assumed to be random.

tics are based on separate analyses of the obtained item scores falling within the same domain and having the same number of digits in their multiplicands, incorporating the design *s:c* × *i*). For most combinations of levels of *d* and *n*, both $\hat{\sigma}^2(i)$ and $\hat{\sigma}^2(i)/\hat{\sigma}^2_{total}$ are quite small. This suggests that for most of the domains considered, the criterion of equivalent difficulty of items within domains is closely approximated. One possible exception to the above conclusion may be found for the items in Domain 15, where the criterion of equal item difficulty is less closely approximated. This may suggest the desirability of exploring the effects of restructuring Domain 15 into a number of more content-restricted subdomains or a recombination of subsections of the domain with other domains. Such modifications are not considered for the data set used in this paper, since the appropriate level of restructuring of a domain-referenced testing system is not at the item sample level but at the item

Table 3
Absolute and Relative Variance Attributable to Items for Specified Combinations of d and n

Digits in Multiplicand and Estimate*	Domain					
	10	12	13	15	17	18
3 Digits						
$\hat{\sigma}^2$(i)	.000	.001	-.000[a]	.005	.002	.001
$\hat{\sigma}^2_{total}$	.213	.202	.236	.250	.247	.247
$\hat{\sigma}^2$(i)/$\hat{\sigma}^2_{total}$	.002	.006	.000	.019	.006	.006
4 Digits						
$\hat{\sigma}^2$(i)	-.000[a]	.001	.001	.003	.001	.001
$\hat{\sigma}^2_{total}$	.207	.219	.242	.236	.250	.236
$\hat{\sigma}^2$(i)/$\hat{\sigma}^2_{total}$	.000	.004	.005	.013	.006	.004

* $\hat{\sigma}^2_{total} = \Sigma\hat{\sigma}^2(\ )$ for the s × i design analyses.

[a]This variance component was assumed to be equal to zero.

universe level, namely, that of complete domains.

The remaining sources of variation presented in Table 2 which accounted for small portions of the total item variance were composed of classroom interaction effects. One possible conclusion to be drawn from these relatively small variances is that this domain-referenced test appears to provide a comparable structure for students from different classrooms.

Comparisons among Coefficients of Generalizability

In an attempt to describe the generalizability that is possible from randomly sampled items from a domain to the item universe from which they were sampled, two different classes of generalizability coefficients related to each domain were computed and compared. These two classes of coefficients related to each domain are presented in Table 4. Note that all generalizability coefficients presented in Table 4 have been corrected for length to a single item or observation test and thus provide estimates of the mean correlations between items within the specified domains (see Cronbach, 1951). These corrections also allow for direct comparisons among all of the coefficients, even though they may be based on different numbers of items.

The coefficients within the first of these two classes were based on the original unconstrained definitions of item domains and thus are equivalent to alpha coefficients corrected to item length one, based on the design $s \times i$ where $n_i = 10$ for each domain and may be defined as

$$\hat{\rho}_u = \frac{\hat{\sigma}^2(s)}{\hat{\sigma}^2(s) + \hat{\sigma}^2(s \times i,e)} . \qquad [1]$$

The second class of coefficients was based on the design $s \times r$, where r represents repeated presentations of an item sampled at two levels ($n_r = 2$). This design provides the maximum possible

constrained conception of domain scope, where only an item and its repeated presentations are considered to be grouped within a set. Thus, the coefficients based on this constrained domain conception are equivalent to alpha coefficients corrected for length as above, based on the replicates of the single item from each domain that was administered on two occasions to examinees and may be defined as

$$\hat{\rho}_c = \frac{\hat{\sigma}^2(s)}{\hat{\sigma}^2(s) + \hat{\sigma}^2(s \times r, e)} . \qquad [2]$$

Under this conception of domain scope, it may be assumed that a logical upper bound for item homogeneity and comparability of item difficulties for domains is obtained. For this reason comparisons between these maximally constrained domains with unconstrained domains may be useful for identifying appropriate content scope for domains. In a similar manner, collapsing all domains into a single set may provide a logical lower bound for item homogeneity and variability of item difficulties.

For each of the domains, only minor to moderate differences in magnitude among the generalizability coefficients were obtained under the two levels of domain constraint, with the coefficients obtained under the constrained approach resulting in slightly larger values. However, all of these coefficients are quite large for a test with a single item and a single observation. Thus, for all domains, very large proportions of total variability in domain scores can be accounted for by small numbers of test items.

To help assess the differences in magnitude found between the pair of generalizability coefficients for each domain, ratios of signal/noise values, $(S/N_c)/(S/N_u)$, related to corresponding pairs of constrained (c) and unconstrained (u) generalizability coefficients corrected to correspond to tests of equal length, are presented in Table 4. These ratios are defined as follows:

Table 4
Assessment of Internal Consistency Among Items Within Domains

Generalizability Coefficient[a]	Domain 10	12	13	15	17	18
$\hat{\rho}_u$	0.551	0.578	0.440	0.338	0.522	0.415
$\hat{\rho}_c$	0.606	0.599	0.443	0.428	0.593	0.576
Ratio of S/N values[b]						
$\frac{S/N_c}{S/N_u}$	1.25	1.09	1.01	1.46	1.33	1.91
$\frac{S/N_u}{S/N_o}$	1.86	2.07	1.19	.77	1.66	1.08

[a]These coefficients have been corrected to correspond to tests containing a single item or observation.

[b]The ratios between S/N values are based on generalizability coefficients corrected to correspond to tests of equal length.

$$\frac{S/N_c}{S/N_u} = \frac{\hat{\sigma}^2(s)_c/\hat{\sigma}^2(s \times r,e)_c}{\hat{\sigma}^2(s)_u/\hat{\sigma}^2(s \times i,e)_u}$$

$$= \frac{\hat{\rho}_c/(1-\hat{\rho}_c)}{\hat{\rho}_u/(1-\hat{\rho}_u)} \qquad [3]$$

where subscripts "*c*" and "*u*," respectively, designate estimated variance components (or generalizability coefficients) based on the constrained and the unconstrained design analyses. Note that the S/N value corresponding to a specific generalizability coefficient is simply that coefficient divided by one minus the coefficient. Such ratios of S/N values specify the proportional test length necessary for the generalizability coefficient related to the unconstrained design (i.e., the generalizability coefficient found in the denominator of the ratio) to equal the constrained design generalizability coefficient at its original test length (see Cronbach & Gleser, 1964). Cronbach, Schonemann, and McKie (1965) have suggested a rule of thumb for assessing such ratios: two coefficients manifest comparable levels of generalizability if the ratio of their S/N values falls between the limits of .83 and 1.20.

The ratios of S/N values based on the two classes of generalizability coefficients presented in Table 4 show considerable variability in magnitude from one domain to another. In the case of Domains 15 and 18 the magnitudes of the above ratios of S/N values suggest that these domains show considerably less internal consistency than would be obtained under maximum content restriction of the domains. In fact, the degree of internal consistency found within these two unconstrained domains is at best no higher than that found among all items across all domains (this latter test structure corresponds to an "overall test" design, $s \times i$, for which the six domains are collapsed and thus $n_i = 60$). This conclusion is supported by the ratio of S/N values (which have been corrected to correspond to tests of equal length) based on the unconstrained and the overall test designs, presented in the last row of Table 4. However, this is not the case for the other four domains. These ratios of S/N values are defined as

$$\frac{S/N_u}{S/N_o} = \frac{\hat{\sigma}^2(s)_u/\hat{\sigma}^2(s \times i,e)_u}{\hat{\sigma}^2(s)_o/\hat{\sigma}^2(s \times i,e)_o} \qquad [4]$$

where the subscript "*o*" designates estimated variance components based on the overall test design analysis.

The implication of these findings for Domains 15 and 18 is that their underlying item form rules are in need of modification. The kind of modifications that would seem to be appropriate are those that would reduce the scope of the content found within the domains. Such modifications of item form rules would, it is hoped, allow the degree of internal consistency found within the restructured domains to better approach their "upper bounds." This, in turn, would allow for better generalizability to domain scores.

Discussion

The results of this example indicate that for the content area considered, the logical approach used to define domains provides at least a reasonable first approximation to desired item groupings. For some of the domains studied, it would appear to be possible to obtain accurate estimation of how students can be expected to perform on an entire domain of items based on their performance on a small sample of items. These domains of internally consistent items provide evidence in support of the contention that greatly overlapping sets of skills are necessary to correctly answer items within such domains.

Additional research is needed to determine whether attempted modifications of "inadequate" item form rules can improve the degree of internal consistency found among items within revised domains. If increased internal consistency is not readily obtainable, this may present an important limitation to the domain-referenced approach for use in diagnostic testing.

References

Brennan, R. L. *Extensions of generalizability theory to domain-referenced testing* (ACT Technical Bulletin No. 30). Iowa City IA: American College Testing Program, June 1978.

Brennan, R. L. *Some applications of generalizability theory to the dependability of domain-referenced tests* (ACT Technical Bulletin No. 32). Iowa City IA: American College Testing Program, April 1979.

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. *Journal of Educational Measurement*, 1977, *14*, 277–289. (a)

Brennan, R. L., & Kane, M. T. Signal/noise ratios for domain-referenced tests. *Psychometrika*, 1977, *42*, 609–625. (b)

Cardinet, J., Tourneur, Y., & Allal, L. The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, 1976, *13*, 119–135.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, *16*, 297–334.

Cronbach, L. J., & Gleser, G. C. The signal-noise ratio in the comparison of reliability coefficients. *Educational and Psychological Measurement*, 1964, *23*, 467–480.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurement: Theory of generalizability for single and multiple observations*. New York: Wiley, 1972.

Cronbach, L. J., Schonemann, P., & McKie, D. Alpha coefficients for stratified parallel tests. *Educational and Psychological Measurement*, 1965, *25*, 291–312.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation No. 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974.

Hartke, A. R. The use of latent partition analysis to identify homogeneity of an item population. *Journal of Educational Measurement*, 1978, *15*, 43–47.

Hively, W. *Domain-referenced achievement testing: Theory and practice*. Unpublished manuscript, 1970.

Hively, W. Introduction to domain-referenced testing. In W. Hively (Ed.), *Domain-referenced testing*. Englewood Cliffs NJ: Educational Technology Publication, 1974.

Hively, W., Patterson, H. L., & Page, L. A "universe-defined" system of arithmetic achievement test forms. *Journal of Educational Measurement*, 1968, *5*, 275–290.

Kane, M. T. A sampling model for validity. *Applied Psychological Measurement*, 1982, *6*, 125–160.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 1947, *61* (4, Whole No. 285).

Lumsden, J. Test theory. *Annual Review of Psychology*, 1976, *27*, 254–280.

Macready, G. B., & Merwin, J. C. Homogeneity within item forms in domain-referenced testing. *Educational and Psychological Measurement*, 1973, *33*, 351–360.

Patterson, H. L., & Vierling, J. S. *EDINET Instruction Series: Individualized Mathematics Program*. Minneapolis MN: Honeywell Information Systems, 1970.

Searle, S. R. *Linear models*. New York: Wiley, 1971.

Author's Address

Send requests for reprints or further information to George Macready, Department of Measurement, Statistics and Evaluation, College of Education, University of Maryland, College Park MD 20742, U.S.A.