A DOMAIN-REFERENCED APPROACH TO DIAGNOSTIC TESTING USING GENERALIZABILITY THEORY

Joan Herman
Noreen Webb
and
Beverly Cabello

CSE Report No. 245
1985

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was in part performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

	Page
Introduction	1
Domain-specification and test development	2
Determining categories for the profile	3
Number of items needed for each score in the profile	4
Estimation of the profile	5
Development and Administration of the Test	7
Design of the Test	7
Test Administration	7
Analyses and Results	8
Generalizability Study of Test Structure	8
Number of Items Per Category in the Profile	10
Estimation of the Profile	10
Conclusions	11
References	14
Footnotes	16
Author Notes	17

INTRODUCTION

Assessment plays an integral role in the improvement of instructional practice. Mastery learning strategies (Bloom, 1976; Block, 1971), systematic instruction (Popham and Baker, 1976), individualized instruction (Glaser, 1970; Klausmeier 1976), clinical teaching (Hunter, 1983), and effective schooling (Edmonds, 1981) all point to the importance of assessment in diagnosing students' strengths and weaknesses, in monitoring their progress through the curriculum, in providing instruction that is tailored to instructional needs and goals and thus in enhancing student achievement.

However, while diagnosis is a recurring concern throughout the instructional process and is central to its success, the assessment tools teachers have available are really quite limited, particularly in language arts. A typical diagnostic test in reading, for example, may provide a total score and subscores for individuals and groups in such areas as vocabulary, literal comprehension, and inferential comprehension, but such scores offer teachers little guidance regarding the nature of any reading problems or their causes. It is left to the teacher to pinpoint why students perform as they do and to prescribe instruction accordingly. Recent research has taken a more molecular view of the diagnostic problem. Tatsuoka and colleagues (Tatsuoka, 1983; Tatsuoka & Tatsuoka, 1983; Tatsuoka et al., 1980), for example, have identified the specific misconceptions and difficulties that students exhibit, but only in mathematics. Their work tends to focus on narrow domains (e.g., the subtraction of two-digit signed numbers), however, and may not be useful for diagnosis in wider subject-matter domains.

The current study presents an intermediate approach for constructing and analyzing useful diagnostic tests for teachers. It combines a domain referenced approach (Hively et al, 1973; Baker, 1974; Popham 1980) with generalizability theory (Brennan, 1983; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1981) to determine factors which may be diagnostically useful in characterizing or profiling students' performance across a range of content areas or domains. The result is a four-stage approach to designing diagnostic profiles for classroom use including: (1) specifying the domain of interest in terms of its constituent factors and developing a test to represent that domain, (2) determining which factor scores should be presented in a profile, (3) determining how many items are needed for dependable measurement of each score, and (4) calculating accurate estimates of the scores in the profile. The approach is illustrated with a specially developed test of pronoun use.

Domain-specification and test development. A domain referenced approach to test design starts with the assumption that the major purpose of testing is to assess an individual's status with respect to a skill or knowledge domain and that valid assessment of that status requires a thorough description of the domain to be assessed. Encapsulated within a domain specification, this description details the content and response parameters which define the domain, providing both a blueprint for generating items and a target for instructional planning (See Baker and Herman, 1983).

In addition to identifying relevant parameters which must be considered to assure that a test provides a representative picture of a skill, domain specifications for diagnostic tests present the additional problem of isolating factors which influence variations in student performance and which predict varying levels of skill proficiency. Items representing these factors can then be appropriately sampled to produce a test with diagnostic utility, i.e., one which identifies the causes or reasons for a students' performance level.

The domain specifications developed for this study were based on factors known to influence learning, particularly those which distinguish expert and novice performance (Chi and Glaser, 1980). The resulting tests represented multiple content factors and levels of cognitive complexity. (For a more detailed description of the domain specification and test development for this study, please see Herman, Webb, & Cabello, 1985.)

Determining categories for the profile. After a test fully covering the domain of interest (and its constituent factors) has been developed and administered, the second step is to determine how many of the initial factors must be preserved in a student's profile. For example, to represent the domain of pronoun usage, the pronoun test developed here initially had 32 different kinds of items, representing 32 separate categories. The purpose of the analysis at this stage, then, was to determine which of these 32 categories needed to be differentiated to adequately represent a student's performance.

This stage uses generalizability (G) theory (Brennan, 1983; Cronbach et al., 1972; Webb & Shavelson, 1981), a measurement theory designed to assess multiple sources of variation in a measurement. In generalizability theory, a measurement is a sample from a universe of admissible observations, characterized by one or more sources of error variation or facets. This universe is typically defined as all combinations of the levels (called conditions) of the facets. An observed score is decomposed into the universe score (analogous to the true score in classical test theory) and error scores corresponding to the multiple sources of error variation. G theory uses analysis of variance to partition sources of variation and estimate the variance component for each source of variation. For example, in a test of pronoun usage with items measuring multiple rules (e.g., nominative, objective pronouns), interest at this stage lies in the variation associated with pronoun rule. If this variation is small compared to the variation between students, then multiple rules do not need to be distinguished in a profile. If the variation due to pronoun rules is large, then each rule must be represented separately in the profile.

Number of items needed for each score in the profile. Once the categories to compose the profile have been determined, the next step is to determine the number of items needed for dependable measurement of each category. Each score entails a generalizability analysis with students crossed with items. Because interest is in absolute decisions, that is, in estimating a student's universe score rather than only comparing him or her to a peer group, the estimate of error variance is

$$\hat{\sigma}^2_{Abs} = \frac{\hat{\sigma}^2_i}{n_i'} + \frac{\hat{\sigma}^2_{pi,e}}{n_i'}$$

where n_i' is the number of items in the decision study, and the generalizability coefficient is

$$\hat{\rho}^2_{Abs} = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \hat{\sigma}^2_{Abs}}$$

where $\hat{\sigma}^2_p$ is the variance component for universe scores.

Estimation of the profile. Although the profile can report observed scores for the categories (here, the observed mean over the items corresponding to a score in the profile), point estimates of the universe scores are more accurate. Two kinds of point estimates are used here: univariate and multivariate. From Cronbach et al. (1972, p. 106), the univariate regression equation for estimating a universe score from the observed score for that variable (univariate estimation) is

$$\hat{\mu}_p = \mu + \rho^2 (X_{pI} - \mu)$$

where X_{pI} is the observed mean for person p over all items corresponding to that score in the profile, μ is estimated from the mean of the sample, and ρ^2 is the generalizability coefficient estimated using equation 2.

Multivariate estimation uses information in the entire profile to estimate the universe score for each category. The multivariate estimate of the universe score for the mth component is

$$\hat{m\mu}_p = \sum_{v=1}^{n_v} \beta_{mv} ({}_vX_{pI} - {}_v\mu) + {}_m\mu$$

where β stands for a multiple regression coefficient and n_v is the number of components in the profile (see Cronbach et al., 1972, p. 313). The multivariate estimates of universe scores generally will be superior to the univariate estimates (Cronbach et al., 1972, p. 313).

The estimation of the multiple regression coefficients requires covariances between categories as well as the universe score variances for each category from the univariate generalizability analyses. The covariances needed are the observed covariances and the covariances between universe scores and observed scores. Because universe and error scores are uncorrelated, the covariances between universe and observed scores are universe score covariances.

Just as the observed score variance can be decomposed into universe score variance and error variances, an observed score covariance can be decomposed into universe score covariance and error covariances (see Webb, Shavelson, & Maddahian, 1983). For example, in the current study, in which the design is persons crossed with items, the observed covariance between two categories can be decomposed into the universe score covariance, the covariance for items, and covariance for the person-item interaction (plus residual error). Because items are independently sampled from each category and because it is assumed that residual effects are uncorrelated with other effects in the model, the error covariances are zero (Brennan, 1983). The universe score covariance, then, is equal to the covariance between observed scores.[1]

DEVELOPMENT AND ADMINISTRATION OF THE TEST

Design of the Test

Review of language curricula and texts, and consultation with content experts showed that a test representing the domain of pronoun usage needed to encompass four content factors and one cognitive complexity factor: pronoun rule (nominative, 3 kinds of objective, possessive), form (relative form -- who or whom -- and non-relative form), number (singular, plural), person (first, second, third), and the level of embeddedness (non-embedded items -- a single sentence -- and embedded items in a paragraph).

The present paper used a subset of the factors to produce a balanced design. The factors in the design were rule (nominative, direct object, indirect object, object of preposition) x form (relative, non-relative) x number (singular, plural) x embeddedness (sentence, paragraph). There were two items for each of the 32 cells in this design. Thus, the portion of the test analyzed here had 64 items. Each item had a multiple choice format with five responses per item.

Test Administration

The test was administered to 128 sixth-grade students from three elementary schools within a local inner-city district. These schools are located in a low to middle SES area with a high transiency rate and a mixed population. Approximately 90 percent of the students were of Hispanic background, 6% were Black, 2% were Asian, and 2% were non-minority Whites.

ANALYSES AND RESULTS

Generalizability Study of Test Structure

In the first generalizability analysis, the object of measurement was persons (students), and the facets were pronoun rule, form, number, embeddedness, and item. All sources of variation were crossed except item, which was nested within the other facets. The items can be considered a random sample from an infinite universe of items and are exchangeable with other items from the universe, hence the item facet is random. Because pronoun rule, form, number, and embeddedness have fixed sets of conditions, they are fixed facets. Following Shavelson and Webb (1981), we chose first to examine the variability of the conditions of these facets, that is, to compute the variance components as if the facets were random. Then, where the variability was substantial, the conditions had to be treated as separate scores in the profile. Where the variability was minimal, the scores would be averaged over conditions of the facet.

The generalizability analysis of the 64-item design yielded 33 variance components. Rather than presenting all 33 variance components, Table 1 presents the variance components for all major sources of variation in the design. Most of the excluded variance components accounted for less than 1% of the total variation and the largest excluded component accounted for 1.6%.

As can be seen in Table 1, three facets -- the form of the pronoun, level of embeddedness, and pronoun rule -- had major effects on performance. However, not all facets were associated with individual differences between students. The effects that did not interact with persons were constant across all students (main effect for Form, Embeddedness-Form interaction, Rule-Form interaction). Considering the Embeddedness-Form interaction, for example, embeddedness had a significant impact on performance for non-relative pronouns, with embedded pronouns (mean proportion correct = .49) being more difficult than non-embedded pronouns (.79) but of similar difficulty for relative pronouns (.29 for embedded, .32 for non-embedded). Because this effect was constant for all students, the teacher needs this information only in a group profile, not in individual students' profiles.

The interactions with persons are associated with individual differences between students. Hence, they show the categories that need to be represented on individual students' profiles. For example, the substantial interaction between person and form shows that some students found relative pronouns to be much more difficult than non-relative pronouns, whereas other students performed nearly the same on both (the interaction was ordinal; no student performed better on relative pronouns than on non-relative pronouns). The major effects that interact with students are rule and form (person-rule-rule-form interaction), so each combination of rule and form must be represented on individual profiles. This results in eight categories:

4 rules x 2 forms. The 32 categories of pronoun usage on the initial test have been pared down to eight categories for the individual student profiles.

Number of Items Per Category in the Profile

On the initial 64-item test, 8 items corresponded to each combination of pronoun rule and form, so a maximum of 8 items were available for the generalizability analyses of each category. The generalizability coefficients for each category with 8 items per category appear in Table 2. For some categories, it would be desirable to have more items to increase the generalizability coefficient. For example, 12 non-relative direct object items would result in a generalizability coefficient of .63. In a real testing situation, more items would be written for each category as needed.

Estimation of the Profile

Each student's universe score was estimated using the univariate and multivariate procedures described earlier. The mean absolute differences between observed scores, univariate estimates, and multivariate estimates in Table 2 show substantial differences between the three sets of scores. The multivariate estimates of universe scores tend to differ more from the observed scores than do the univariate estimates. The greater accuracy of the multivariate estimates than the univariate estimates is shown by the R^2 and percent error reduction in Table 2. (The R^2 for the univariate estimates is equal to the generalizability coefficient.) The multivariate estimation results in substantial error reduction, ranging from 14% to 78% across the eight categories.

Figure 1 shows an example profile to illustrate the difference between observed scores, univariate estimates, and multivariate estimates. The profiles of estimated universe scores are flatter than the profile of observed scores. If a teacher used a criterion score of 70% correct to decide whether to provide this student with further instruction, the teacher would err on three categories using the observed scores. The observed scores suggest that this student has mastered four categories, whereas the multivariate profile suggests only two. Interestingly, one of the categories showing mastery using the multivariate profile does not show mastery using the observed scores. The univariate profile shows no mastery of any category using the 70% criterion.

Not only did observed scores, univariate estimates of universe scores, and multivariate estimates of universe scores give different pictures of mastery of pronoun usage, they rank ordered students differently. Whereas observed scores and univariate estimates gave the same rank order of students, the multivariate estimates sometimes changed the rank order. Correlations between observed scores and multivariate estimates of universe scores ranged from .70 to .97 across categories in the profile.

CONCLUSIONS

This paper described a four-step approach to constructing diagnostic test profiles that provide precise but practical information on students' problems and their need for additional

instruction or remediation. The first step was specifying the domain of interest and constructing and administering a test representing that domain. The test of pronoun usage in the current study had 64 items measuring 32 different kinds of pronoun usage. The second step was a generalizability study to determine which of the 32 categories showed substantial variation among students. The results showed that eight of the 32 categories needed to be distinguished in a profile: relative and non-relative forms of nominative pronouns, direct object, indirect object, and object of preposition. The third step was to determine the number of items needed for dependable measurement of each of the eight categories in the profile. The results showed that all eight items corresponding to each category were needed for dependable measurement. The final step was to estimate universe score profiles. Comparison of observed scores, univariate estimates and multivariate estimates of universe scores showed that multivariate estimates were more accurate and often produced different profiles from both univariate estimates of universe scores and observed scores.

While the four-step process to test development described here is far too complex and time-consuming for a classroom teacher to undertake, the resultant test and score profiles are feasible for classroom use, particularly if the analytic, scoring and recordkeeping processes are automated on microcomputers. With a single, relatively short test, a teacher could then correctly identify students' strengths and weaknesses on all important aspects of a curriculum domain and make instructional decisions accordingly.

The approach to diagnostic testing described here involved a single administration of the complete test under one set of conditions. This approach can, however, be extended to include other factors in the testing situation that might be important, for example, format of the test or test items (e.g., multiple choice vs. completion items), vocabulary level of the items, content of the material in the items or occasion of testing. If there is reason to believe that the format of the test may influence test performance, different versions of the initial test can be administered to students and the effect of test format on performance can be estimated. If both test format and occasions are suspected to influence test performance, multiple versions of the test could be administered on multiple occasions, making it possible to estimate the effect of both factors on performance. Moreover, these effects would be estimated simultaneously with estimating the effects of different factors in the test structure in the generalizability analysis (Step 2 in the four-step approach described here). Thus, it would be possible to estimate interactions among all of these factors (for example, performance on some topics in the domain may differ across versions of the test while performance on other topics does not). The power of this approach is great; the possible factors to estimate are limited only be constraints on time (e.g., for testing students) and resources (e.g., for developing different versions of the test).

REFERENCES

Baker, E.L. (1974). Beyond objectives: Domain referenced tests for evaluation and instructional improvement. Education Technology, 14.

Baker, E.L., & Herman, J. (1983). Task structure design: Beyond linkage. Journal of Educational Measurement, 20(2), 149-164.

Block, J.H. (1971). Mastery learning: Theory and practice. New York: Holt, Rinehart and Winston.

Bloom, B.S. (1976). Human characteristics and school learning. New York: McGraw Hill Book Company.

Brennan, R.L. (1983). Elements of Generalizability Theory. Iowa City, Iowa: American College Testing Program Publications.

Chi, T.H., & Glaser, R. (1980). The measurement of expertise: Analysis of the development of knowledge and skill as a basis for assessing achievement. In E.L. Baker and E.S. Quellmalz (Eds.), Educational testing and evaluation. Design, analysis, and policy. Beverly Hills, California: Sage Publications.

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

Edmonds, R. (1981). Making Public Schools Effective. Social Policy, 12, 56-60.

Glaser, R. (1970). Evaluation of instruction and changing educational models. In M.C. Wittrock and D.E. Wiley (Eds.), The evaluation of instruction. Chicago: Rand McNally and Company.

Herman, J., Webb, N., & Cabello, B. (1983). A referenced approach to diagnostic testing. (CSE Technical Report No. 245). Los Angeles, CA: UCLA Center for the Study of Evaluation.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. (1973). Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST project. CSE Monograph No. 1. Los Angeles: Center for the Study of Evaluation, University of California.

Hunter, M. (1983). Mastery teaching: increasing instructional effectiveness in secondary school, colleges and universities. TIP Publications, El Segundo, California.

Klausmeier, H.J. (1976). Individually guided education: 1966-1980. Journal of Teacher Education, 3, 199-206.

Popham, W.J. (1980). Domain specification strategies. In R.A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore: John Hopkins University Press.

Popham, W.J., & Baker, E.L. (1976). Classroom instructional tactics. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

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

Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20, 345-354.

Tatsuoka, K.K., Birenbaum, M., Tatsuoka, M.M., & Baillie, R. (February 1980). A psychometric approach to error analysis on response patterns. (Research Report 80-3). Urbana, Illinois: University of Illinois, Computer-based Education Research Laboratory.

Tatsuoka, K.K., & Tatsuoka, M.M. (1983). Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement, 20, 221-230.

FOOTNOTES

1. It could be argued that the error covariances are correlated (called linked, see Cronbach et al., 1972) due to the fact that the items were part of the same test administered on one occasion. Without administering the test on multiple occasions, however, it is not possible to estimate this effect.

AUTHOR NOTES

The project presented or reported herein was performed pusuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the policy of the National Institute of Education,and no official endorsement should be inferred.

A version of this paper was presented at the 1985 Annual Meeting of the American Educational Research Association, Chicago.

We would like to thank Richard Shavelson for his helpful comments on an earlier draft of this paper.

Requests for reprints should be sent to Noreen Webb, Graduate School of Education, UCLA, Los Angeles, CA 90024.

TABLE 1

Generalizability Study of Test Structure

Source of Variation		Percent of Total Variation
Persons (P)	.02000	6.4
Pronoun Form (F)	.04128	13.3
F x Embeddness (E)	.01455	4.7
F x Rule (R)	.02300	7.4
P x R x F	.01801	5.8
Residual	.14520	46.7
All other	.04911	15.7

TABLE 2

Results of Generalizability Analyses for
Each Category in Profile

Pronoun Form and Rule	Mean Observed Score [a] for 8 items	Mean Absolute Difference [a]			R^2 Multivariate	Percent Error Reduction From Univariate to Multivariate
		Observed Score vs Univariate Estimate	Observed Score vs Multivariate Estimate	Univariate Estimate vs Multivariate Estimate		
Non-Relative						
Nominative	.55	.11	.11	.07	.84	75
Direct Object	.59	.08	.11	.08	.84	66
Indirect Object	.68	.06	.09	.07	.87	55
Object of Preposition	.70	.07	.09	.07	.82	45
Relative						
Nominative	.57	.09	.09	.03	.60	9
Direct Object	.18	.07	.08	.06	.87	70
Indirect Object	.22	.07	.08	.05	.75	67
Object of Preposition	.25	.05	.09	.06	.87	48

[a] Proportion Correct