OPTIMIZING THE DIAGNOSTIC POWER OF TESTS: AN ILLUSTRATION FROM LANGUAGE ARTS

Noreen Webb
Joan Herman
Beverly Cabello

CSE Report No. 226
1984

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

The project presented or reported herein was 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

INTRODUCTION

Assessment has an integral role to play 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. The underlying theory derives from a systems view of education and suggests that if teachers are to maximize their students' learning, they need to: plan instruction on the basis of the needs of individual or groups of students; monitor their progress; determine whether remediation is required; and evaluate outcomes to assess the success of instruction, as well as needs for modification and students' readiness for succeeding work.

Diagnosis and prescription is thus a recurring concern throughout the instructional process and is central to its success. Yet despite its importance, the assessment tools teachers have available for such a process are really quite limited. While so-called diagnostic tests do exist, the level of information they provide is less than optimal. A typical diagnostic test in reading, for example, may characterize student needs by providing a total score and subscores for individuals and groups in such areas as vocabulary, literal comprehension, inferential comprehension, etc, 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. In contrast to this global approach, more recent research has taken a molecular view of the diagnostic problem. Tatsuoka and associates (1980) for example, have completed extensive work in diagnosing student performance in a very narrow mathematics domain (the subtraction of two digit signed numbers) and have identified the specific misconceptions and difficulties which students manifest in this area, e.g., six specific error types related to determining the sign of answers. While the advent of classroom computer technology may make such advances more useable in future classroom practice, these findings provide a level of detail beyond the grasp of today's teachers: a teacher cannot track a classroom of students across so many error dimensions for all curriculum areas, nor feasibly tailor instruction at this level of specificity.

The current study seeks an intermediate level for constructing and analyzing diagnostic tests for classroom use. It investigates strategies for improving the power of diagnostic instruments so that they provide more precise but practical information on students' problems and needs. Based on a domain referenced approach (Hively et al, 1973; Baker, 1974; Popham 1980), the study examines factors which may be diagnostically useful in characterizing or profiling students' performance across a range of content areas or domains; i.e., factors which may be used to structure the test domain, which predict and conceptually define item difficulty, and which likewise may be used to

structure instructional treatments. The study also explores methods for analyzing and structuring diagnostic tests so that the process is efficient and does not place an undue information load on teachers. Specifically the study addresses three inter-related questions:

1. What factors ought to be considered in specifying a test domain so that the resultant test will provide specific instructionally relevant diagnostic information?
2. What analysis procedures can be used to optimally structure a diagnostic test to provide valid, reliable, and efficient profiles of individual and group performance?
3. Are the subject strategies feasible for classroom use, or do they require an unrealistic investment in time or an impractical level of detail?

In order to address these questions, we developed an illustrative test of pronoun use representing factors of interest; administered the test to a heterogeneous sample of sixth grade students; used generalizability theory to analyze results and suggest an optimal test structure; and reflected on the entire process to assess feasibility and implications for practice. In the sections which follow, we first describe the domain referenced framework which guided the test development process and the specific factors which were chosen for scrutiny, followed by a description of the test, the analytical approach, and our conclusions.

A DOMAIN REFERENCED APPROACH TO TEST DESIGN

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 understanding and specification of the domain to be assessed. The objective of assessment, in other words, needs to be well defined to assure that a test actually measures what it is intended to measure and that items reflect test content. The definition is reflected in a domain specification which provides a blueprint for developing test items and can serve also to target effective instructional sequences.

While a number of approaches to domain specification have been proposed, all seek to define a pool of items that represents an important universe of knowledge or skill domain such that student performance in one set of items drawn from the domain would generalize to a second set of items and to the entire domain. In its most highly prescribed form, domain specifications provide an exhaustive set of rules for generating a set of test items (Hively et al, 1973; Osburn, 1968; Millman, 1980). As more commonly practiced (and as exemplified in the present study) domain specifications provide a conceptual map of the skill to be assessed, including relevant parameters for defining the range of eligible content, the response level to be represented in the item, item format, directions, and a sample item (Baker; 1974, Popham, 1980; Hambleton, 1980).

Regardless of approach, the identification of relevant parameters becomes a central problem. Establishing content limits is an initial concern, most commonly solved by reference to extant curricular material, subject area specialists and/or mutually agreed upon goals and boundaries, -- or more preferably research on the structure of the

knowledge base and the nature of learning and development. Establishing response limits, including criteria for judging constructed responses and rules for generating incorrect alternatives in selected responses, fixes attention on the quality of expected performance, the level of response differentiation desired, and systematic error patterns that may be operable. Framed by linguistic complexity, form of content, and cognitive complexity, test content is specified to represent the domain of interest (see Baker and Herman, 1983).

In addition to identifying content parameters which must be included to assure that a test provides a representative picture of a skill, diagnostic tests present the additional problem of isolating factors which influence variations in student performance and predict varying levels of skill proficiency. In other words, it is critical to identify important factors within a domain which cause an item to be more or less difficult or a student's performance to vary. 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 students' performance level.

What variables might be useful for constructing such diagnostic profiles? Research in cognitive psychology provides some clues. Chi and Glaser (1980) propose a framework for understanding the nature of differences between expert and novice performance. Their framework characterizes information processing in terms of two components: knowledge or content structure, and cognitive processes, components which are well supported in the research literature. Various authors,

for example, have pointed to the effect of cognitive processing demands inherent in a task. Principal distinctions have been made for tasks which require storage, association, and retrieval of information contrasted with tasks requiring processing of information, including subordination, reconfiguration, and other adaptive processes (Spiro, 1980; Quellmalz, 1982).

Beyond their theoretical justification, content structure and cognitive complexity are appealing also in terms of their feasibility for practical use. Teachers of course are well used to dealing with the structure of content (at least as their curriculum or instructional materials define it), and their coverage of that content, as research (and intuitive logic) amply demonstrates, is strongly related to student test performance. Cognitive complexity, while perhaps not in the common parlance of classroom teachers, can be operationally defined to be easily accessible to them.

The present study is derived from the foregoing framework of domain referenced testing. A diagnostic test was developed to assess one skill within the language arts curriculum. The domain and item pool were developed to assess the effects on student performance of content structure and cognitive complexity to examine their utility for constructing diagnostic profiles. The test development process is described in the section which follows.

DEVELOPMENT AND ADMINISTRATION OF THE TEST

Design of the Test

After selecting language arts as a target area for test development, local teachers and administrators were asked to indicate

the kinds of grammar problems their students most frequently exhibited at the upper elementary and junior high school grade levels. One of the most common responses was that students have difficulty with pronouns, particularly in identifying the correct pronoun referent or in using pronouns correctly. Teachers also indicated that a diagnostic test of pronoun use would be beneficial for their classroom instruction. Pronoun use was therefore selected as an appropriate topic for diagnostic test development.

Following the procedures outlined above, language curricula, texts, and content experts were consulted to specify the test domain. Specifications of linguistic properties, e.g., the recurrence and complexity and sequencing of the vocabulary and phrases, were also included to assure that the language would be clear and comprehensible to the test taker and that the test would therefore be a measure of pronoun use rather than reading comprehension, (see, for example, Doehring and Aulls, 1979). Distractor rules were developed systematically to reflect common usage errors. The domain specification reflected in particular the two factors selected for inquiry.

The content structure factor. The curricular review showed that nominative, objective, (including direct object, indirect object, and object of the preposition) and possessive pronouns appear most frequently. These five types of pronouns (including the three objective forms) correspond to rules of grammar, and are called pronoun rules in this paper. The review further revealed that the pronouns corresponding to each rule can also be classified by form,

number, and person. There are two types of form: relative form (who or whom) and non-relative form. Number pertains to singular (she) and plural (they). Person can be of three types: first (I, we), second (you), and third (he, she, they). Since items measuring the second person would have sounded contrived to the reader, the test developed here included only the first and third persons.

The cognitive complexity factor. The two levels of cognitive complexity corresponded to whether students had to use the context of a reading passage to determine the correct pronoun. In the first level, the student was presented with a single sentence that included an underlined noun(s). The student was to select the pronoun to match the underlined noun(s). In other words, the pronoun referent was given and the student needed only to associate that referent with the correct pronoun. In the second, more complex level, students were presented with a short paragraph that included a blank in the place of one noun; students needed to use the context of the paragraph to identify the referent that was appropriate to the blank and then select the correct pronoun for that referent. The correct pronoun could be determined only from elements of the paragraph in which the pronoun was embedded. Consequently, the test developed here used two levels of embeddedness corresponding to two levels of cognitive complexity; non-embedded items (a single sentence) and embedded items (a paragraph).

In summary, the test had five pronoun factors including four representing content structure and one representing cognitive complexity: pronoun rule (nominative, three types of objective,

possessive), pronoun form (relative, non-relative), pronoun number (singular, plural), pronoun person (first, third), and embeddedness of the pronoun (single sentence, paragraph).

Structure of the test. To investigate the impact of each factor on test performance, items were generated for as many combinations of the factors as possible. For each combination, two parallel items were written. The ideal test would have items for every combination of the five factors. Since the form, embeddedness, person, and number factors each had two levels and the rule factor had five levels, a complete test would have 80 (2 X 2 X 2 X 2 X 5) combinations. However, for several combinations of factors, sensible items could not be written. First, non-embedded items could not be written to elicit singular first person pronouns (I, me, or my). Second, items testing the relative form of first-person pronouns would have been contrived. Third, there exist no relative form of possessive pronouns. Excluding these 34 combinations of factors leaves 46 combinations. Since two parallel items were written for each combination, the total test had 92 items. The total design of the test is presented in Table 1.

The analytic approach used here to analyze the test structure requires a fully crossed, balanced design. Since the design of the total test was unbalanced--34 cells in Table 1 are empty--it was necessary to divide the total design into three fully crossed, balanced designs to represent all cells in the design. Design I represented the combination of five factors: for (2 levels), embeddedness (2 levels), rule (4 levels), number (2 levels), and items (2 levels). This design had 64 items. As indicated in Table 1, the

Table 1
Design of the Pronoun Test

	Non Relative Pronoun								Relative Pronoun							
	Non-Embedded				Embedded				Non-Embedded				Embedded			
	1st Person		3rd Person		1st Person		3rd Person		1st Person		3rd Person		1st Person		3rd Person	
Rule:	Sing.	Plur.	Sing.	Plur.	Sing.	Plur.	Sing.	Plur.	Sing.	Plur.	Sing.	Plur.	Sing.	Plur.	Sing.	Plur.
Nominative	—[a]	2	2	2	—	2	2	2	—	—	2	2	—	—	2	2
Direct Object	—	2	2	2	—	2	2	2	—	—	2	2	—	—	2	2
Indirect Object	—	2	2	2	—	2	2	2	—	—	2	2	—	—	2	2
Object of Preposition	—	2	2	2	—	2	2	2	—	—	2	2	—	—	2	2
Possessive	—	2	2	2	—	2	2	2	—	—	—	—	—	—	—	—

[a] No items in this cell.
[b] 2 items per cell.

inclusion of the form factor made it impossible to include items measuring first person pronouns and items measuring possessive pronouns.

The two remaining designs were formed to include the possessive rule. Since the possessive rule applies only to non-relative pronouns, these two designs consisted only of non-relative items. One design (Design II) incorporated the contrast between singular and plural pronouns (number). The other design (Design III) incorporated the constrast between first person and third person pronouns (person). Design II, then, included four factors: embeddedness (2 levels), rule (5 levels), number (2 levels), and item (2 levels), resulting in 40 items. Design III also included four factors: embeddedness (2 levels), rule (5 levels), person (2 levels), and item (2 levels), resulting in 40 items. Many items in the test were included in more than one of the three designs. All of the analyses presented in this paper focus on these three designs.

Structure of the item. The test used a multiple choice format with five alternatives per item, consisting of the correct response and three distractors which were correct in all ways but one, and a fourth distractor which was correct in only one way or not at all. An example is the following item, "Mom praised Mary and Stevie", with the following alternatives: them, they, us, him and she. The correct response (them) is an objective, plural third-person pronoun. The next three responses (they, us, and him) were correct on two of the three factors (rule, number, person). The final response (she) was correct only in the person. The last response was considered a "wild card" distractor (a highly unlikely selection). Such distractors were

included to detect guessing or carelessness.

Test Administration

Through pilot administrations and feedback from teachers and students, the test was modified three times. The final diagnostic 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 rate of transience 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. There were 79 students classified as FEP (Fluent English Proficient) and 49 classified as LEP (Limited English Proficient). Language classification was indicated by the district, based on district reclassification criteria of language proficiency tests, achievement tests, and teacher judgment.

Two forms of the diagnostic test were prepared. Both contained the same items but the order of the items was inverted: items that appeared on the first half of Form A were placed on the second half of Form B and vice versa.

Staff researchers were trained to administer the test. The test instructions allowed the administrators to clarify the meaning in vocabulary item stems but not in item distractors. The tests were administered at the schools. Students were allowed up to 90 minutes to complete the test although most students finished the test in about 45-60 minutes. Classroom teachers were present during testing.

OVERVIEW OF THE ANALYTIC APPROACH

The test score that a teacher uses to evaluate students' grasp of a curricular unit is typically the total score. If the whole class does poorly on a test of fractions, the teacher may decide to spend more time on the unit. If some students in the class do poorly on the test, the teacher may provide them with remedial instruction. Traditional approaches to reliability in educational and psychological measurement concern the dependability of that total score. The approaches focus on the consistency of students' scores over time, (test-retest reliability), from one test form to another (parallel forms reliability), or on the consistency of students' performance across items or sections of a test (internal consistency reliability).

Traditional Approach to Internal Consistency

Of the traditional approaches to reliability, only internal consistency reliability addresses the variability of performance across items within a test. Internal consistency alpha, for example, indicates how consistent student performance is across all items in a test. The magnitude of the coefficient shows whether the rank-ordering of student performance is stable across all items. A high value of alpha (at or near 1.00) indicates that the students who perform better than other students on one item also do so on the other items. A low value of alpha (at or near zero) indicates that the students who perform best on some items are not the same students who perform best on other items. The latter result suggests that all items on the test are not measuring the same construct, and that student performance is different across different parts of the test.

In this situation, the total test score is probably a poor indicator of students' mastery of the material.

While traditional approaches to internal consistency reliability provide some information about the consistency of performance across items in a test, they have limited usefulness for diagnosing specific areas of difficulty. For diagnostic purposes, it is important to have information about student performance on different parts of the test, i.e., a profile of scores. In the test of pronouns developed in the current study, it would be possible to obtain separate scores for each rule of speech (nominative, objective, etc.), for singular and plural items, for first and third person items, and for each form of item (embedded in multiple sentences or non-embedded). While it would be possible to obtain such a detailed profile of scores for each student, this level of detail may not be necessary and might not be worth the cost of obtaining it. The central question is what level of detail in a profile is necessary to inform a teacher about difficulties that individual students or groups of students are having with the material.

The Analytic Questions

The analytic approach used in the present study focuses on the consistency of students' performance across multiple dimensions of a test, each dimension designed to measure a different aspect of the curricular unit. The aim of the analysis is to determine the minimum amount of information about student performance on the test that needs to be presented to guide teachers' future instructional decisions for individual students or for groups of students. The analysis addresses three issues: (1) the necessity of computing profiles of scores for

individual students rather than only one for the class (or one for each subgroup of students in the class), (2) the level of detail that is necessary in the group or individual profiles, and (3) the number of items that are needed to obtain reliable scores in a profile.

Regarding the first issue, if all students have difficulty with the same material (for example, all students misunderstand how to use possessive pronouns), then a single profile for the whole class may be sufficient for diagnosing areas of difficulty. If some material is particularly troublesome to some students but is not troublesome to other students, then profiles for individual students may be necessary. Regarding the second isssue, if students perform equally well on all rules (nominative, objective, possessive), then it would not be necessary to provide separate scores for each rule. If, on the other hand, mastery of nominative pronouns is much greater than that of possessive pronouns, then it would be necessary to include in the profile separate scores for each rule. Regarding the third issue, once it is determined what scores should be included in a profile, the question remains about the number of items that are needed to reliably measure each skill represented in the profile.

Multidimensional Approach to Test Structure: Generalizability Theory

Sketch of generalizability theory. To address the above issues, performance on the pronoun test was analyzed using generalizability theory. Generalizability (G) theory is a measurement theory designed to assess multiple sources of variation in a measurement (see Cronbach, Gleser, Nanda, & Rajarantnam, 1972; Shavelson & Webb, 1981; Webb & Shavelson, 1981). In a nutshell, G theory uses analysis of

variance to partition sources of variation in measures of performance of behavior. The results of a generalizability study show the relative magnitudes of the sources of variation in a test and can be used to improve its design.

A measurement is a sample from a universe of admissible observations, characterized by one or more sources of error variation or facets (e.g., items, rules of grammar). This universe is typically defined as all combinations of the levels (called conditions in G theory) of the facets. Since different measurements may represent different universes, G theory speaks of universe scores rather than true scores, acknowledging that there are different universes to which decision makers may generalize. Likewise, the theory speaks of generalizability coefficients rather than the reliability coefficient, realizing that the value of the coefficient may change as definitions of the universe change.

In G theory, a measurement is decomposed into a component for the universe score and one or more error components. As an illustration, consider a 10-item test of pronoun knowledge in which 5 items measure singular pronouns and 5 items measure plural pronouns. This test has two facets: pronoun number (singular vs. plural) and item. If 20 students take this test, then the design underlying this study is a two-facet partially nested design with items (i) nested within pronoun number (n) and crossed with student (s). The object of measurement, here students, is not a source of error and, therefore, is not a facet.

The variance of the observed scores on this test (over all

students and all items for each pronoun number) can be decomposed into independent sources of variation due to differences between students, items, and pronoun number and the interactions among them using analysis of variance. From the analysis of variance, an estimate of each component of variation in the scores is obtained: $\hat{\sigma}^2_s$, $\hat{\sigma}^2_n$, $\hat{\sigma}^2_{i,ni}$, $\hat{\sigma}^2_{sn}$, and $\hat{\sigma}^2_{si,sni,e}$. (Since items are nested within pronoun number in this design, the main effect for item (i) is confounded with the interaction between item and pronoun number (ni).) G theory focuses on these variance components. The relative magnitudes of the components provide information about particular sources of variation influencing performance on the test. The estimated variance component for students, $\hat{\sigma}^2_s$, is the universe score variance and is analogous to the true score variance in classical theory. The remaining variance components are considered error components.

G theory recognizes that decision makers (teachers, for example) may use the same score in different ways. Some interpretations focus on individual differences (relative decisions). For example, the teacher may be concerned mainly with the generalizability of the rank ordering of students, in order to give remedial instruction to the ten lowest-scoring students. Other interpretations may focus on the level of student performance itself, without reference to other students' peformance (absolute decisions). For example, the teacher may be concerned about a student's absolute level of pronoun knowledge, not how well he or she does relative to other students in the class.

Measurement error is defined differently for each of these proposed interpretations. For relative decisions, the error variance consists of all variance components representing interactions with the object of measurement (here, students):

$$\hat{\sigma}^2_{Rel} = \frac{\hat{\sigma}^2_{sn}}{n_n} + \frac{\hat{\sigma}^2_{si,sni,e}}{n_i n_n}$$

In the above equation, n_n is the number of levels of the pronoun number facet and n_i is the number of items per pronoun number. The error variance for relative decisions reflects differences in rank ordering of students across items and pronoun number. If an interaction effect is large, then students' scores are not rank ordered the same across levels of the facet. For example, if the component representing the interaction between students and number is large relative to the other components, then students who perform the best on singular items are not the same students who perform the best on plural items.

For absolute decisions, the error variance consists of all variance components except that for universe scores:

$$\hat{\sigma}^2_{Abs} = \frac{\hat{\sigma}^2_{n}}{n_n} + \frac{\hat{\sigma}^2_{i,ni}}{n_i n_n} + \frac{\hat{\sigma}^2_{sn}}{n_n} + \frac{\hat{\sigma}^2_{si,sni,e}}{n_i n_n}$$

The error variance for absolute decisions reflects differences in mean performance of students across items and pronoun number as well as differences in rankings of students. When the decision maker is concerned with the absolute level of student performance, the variance components associated with effect of pronoun number and items ($\hat{\sigma}^2_n$ and $\hat{\sigma}^2_{i,ni}$) are included in error variance. The difficulty of one item as compared with another will influence a person's score. A test composed of easy items will suggest a higher level of proficiency than a test composed of difficult items. A large component for pronoun number, as another example, indicates that students find items of one number (say, plural) more difficult than items of the other number (singular).

Generalizability theory and score profiles. The relative magnitudes of the variance components contributing to relative error variance and absolute error variance can be used to determine what kinds of score profiles are necessary for diagnostic purposes. Wherever variance components contributing to relative error variance (interaction with students) are large, separate profiles are necessary for diagnosing learning difficulties. If the interaction between students and pronoun number is large, separate profiles would show which students were having more difficulty with plural items than singular items and which students were having more difficulty with singular items than with plural items. If the variance components contributing to relative error (interactions with students) are small, but the remaining components that contribute to absolute error (components that do not involve interations with students) are large,

then one profile for the class would be sufficient. For example, if all students find plural items more difficult than singular items (a large variance component for pronoun number, $\hat{\sigma}^2_n$), then a profile for the class (the means for singular items and plural items) would show the average difference between plural and singular items. Finally, if the variance components that contribute to relative error variance and absolute error variance are both small, then student performance does not vary across the dimensions of the test. In this case, the total score on the test would be sufficient to guide decisions about instruction.

The above description concerns the relative magnitudes of the variance components, that is, the proportion of total variance accounted for by each variance component. A difficult decision is what proportion is to be considered large. There is no rule of thumb about what proportion should be considered large. In the present study, all variance components that account for at least 3.5 % of the total variation will be noted and discussed. This level is conservative; other researchers might set a level of 5% or even 10% as the minimum proportion that should be used. As in all decision studies, there is a trade-off between cost and efficiency and information. Using a small proportion as a minimum may produce more detailed profiles than are necessary. Using a large proportion as a minimum, on the other hand, may cause important sources of variation to be overlooked or disregarded.

The optimal number of items in a profile. While stressing the importance of variance components and error variances, G theory also

provides a coefficient analogous to the reliability coefficient in classical theory. The generalizability coefficient for relative decisions is defined as:

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

An analoguous coefficient can be defined for absolute decisions:

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

The generalizability coefficient, $\hat{\rho}^2$, indicates the proportion of observed score variance ($\hat{\sigma}^2_s + \hat{\sigma}^2_{Rel}$ or $\hat{\sigma}^2_s + \hat{\sigma}^2_{Abs}$) that is due to universe score variance ($\hat{\sigma}^2_s$). As the number of observations per student increases (for example, the number of items), the error variance ($\hat{\sigma}^2$ or $\hat{\sigma}^2_{Abs}$) decreases and the generalizability coeficient ($\hat{\rho}^2$) increases.

In the present context, the generalizability coefficient is useful for determining the number of items needed to provide a generalizable measure of each score in a profile. If the relative magnitudes of the variance components show that separate scores are needed for each student for plural pronouns and for singular pronouns (indicated by a large interaction between students and pronoun number), then one generalizability analysis would be performed for plural items and another one would be performed for singular items.

The design of each generalizability analysis is simple: student is crossed with item. This design has three variance components: one for students (σ^2_s), one for items (σ^2_i), and one for the interaction between students and items plus unexplained residual variation ($\sigma^2_{si,e}$). The error variance for relative decisions is:

$$\hat{\sigma}^2_{Rel} = \frac{\hat{\sigma}^2_{si}}{n_i}$$

and the error variance for absolute decisions is:

$$\hat{\sigma}^2_{Abs} = \frac{\hat{\sigma}^2_i}{n_i} + \frac{\hat{\sigma}^2_{si}}{n_i}$$

If the analysis shows that a suitable level of generalizability (say, .70) can be obtained with 10 items, then the test would include 10 plural pronoun items and a student's mean on these 10 items would constitute his or her score for plural items in the profile.

If the variance components indicate that a profile of group mean scores is appropriate, then the object of measurement is the group, not the student, and the analysis changes accordingly (see Kane & Brennan, 1977). In the illustration used in the present study, there are two groups of students defined by their language backgroud: fluent English proficient and limited English proficient. In determining the mean score for a language group, the object of measurement is the language group. So the estimated variance

component for language background $(\hat{\sigma}^2_l)$ is the universe score variance. The variation among students is error variation and so student becomes a facet of error. The design of the generalizability analysis of the number of items needed to measure a score in the group mean profile is students (s) nested within language group (l) and crossed with item (i). The error variance for relative decisions is:

$$\hat{\sigma}^2_{Rel} = \frac{\hat{\sigma}^2_{s,sl}}{n_s} + \frac{\hat{\sigma}^2_{si,sli,e}}{n_s n_i}$$

and the error variance for absolute decisions is:

$$\hat{\sigma}^2_{Abs} = \frac{\hat{\sigma}^2_{i}}{n_i} + \frac{\hat{\sigma}^2_{li}}{n_i} + \frac{\hat{\sigma}^2_{s,sl}}{n_s} + \frac{\hat{\sigma}^2_{si,sli,e}}{n_s n_i}$$

If the analysis shows that 10 items are needed to produce a dependable measure of the group's knowledge of plural items, then the test should have 10 plural items.

Summary. In summary, the issues of the appropriate score profiles for diagnostic purposes and the number of items needed to produce dependable measures of each score in the profile are addressed in two stages. The first stage is a generalizability study of the structure of the test. In the illustration presented in this paper, the facets include: rule of grammar, pronoun number, context (embedded vs. non-embedded), person (first person vs. third person, form (relative vs. non-relative), and item. The relative magnitudes

of the variance components in this design show which score should be included in individual student profiles. The second stage is a separate generalizability analysis for each skill in the individual and group profiles to determine the number of items that should be included in the test to obtain a dependable measure of those skills.

RESULTS OF ILLUSTRATIVE ANALYSES

This section illustrates the analytic approach to diagnostic testing described in the previous section. It summarizes (1) the preliminary analyses to determine which population subgroups to include in the generalizability analyses; (2) the three designs underlying the generalizability analyses of test structure; (3) the variance components produced by the generalizability analyses, (4) example diagnostic profiles; and (5) the number of items that would be needed to yield dependable measures of each score in the diagnostic profiles.

Preliminary Analyses

The first step in the approach to diagnostic testing presented here is to determine whether there are distinct population subgroups in the design. In the present illustration, the pronoun test was administered to students from multiple classrooms and schools, and students differed in ethnic background, language background, and age. Therefore, preliminary analyses were conducted to determine whether these factors influenced performance on the pronoun test. Analysis of variance F tests revealed that the only population characteristic influencing performance on the test was language background (FEP vs. LEP; $F(1) = 30.09$, $p < .001$). The statistical tests for classroom,

school, ethnic background, and age were not significant (F statistics ranged from .12, $p < .73$ to 1.06, $p < .37$. In all further analyses, then, only the distinction between FEP and LEP students was maintained.

Summary of the Three Designs

As was described in the section summarizing the design of the test, the entire test can be described by three crossed designs. Design I is a five-facet design yielding 64 items: embeddedness (2 levels), pronoun form (2 levels), rules (4 levels), number (2 levels), and 2 items for each combination of the previous four facets. Design II is a four-facet design yielding 40 items: embeddedness (2 levels), rules (5 levels), number (2 levels), and 2 items for each combination of the previous facets. Design II is also a four-facet design yielding 40 items: embeddedness (2 levels), rules (5 levels), person (2 levels), and 2 items for each combination.

Variance Components and Descriptive Analyses

As a result of the complexity of each design, the number of variance components in each analysis was very large. For example, in the analysis of Design I, with students nested within language background and students and language background crossed with embeddedness, rules of grammar, number, form, and item, there were 51 variance components. Rather than present the descriptive results (means, standard deviations) for all variance components in each design, descriptive results are presented only for components that account for at least 3.5% of the total variation in the design. Table 2 presents the variance components that exceed 3.5% of the total variation in each of the three designs. Each number in Table 2

represents the percentage of total variation accounted for by each variance component. Tables 3, 4, and 5 present the means and standard deviations corresponding to all variance components listed in Table 2. The means are the percent correct, so the maximum score possible is 1.00.

Variation due to student and language background. The large component for language background in each design indicated that FEP and LEP students showed different levels of performance on the test. As the descriptive results in Tables 3 through 5 show, FEP students showed higher mastery of pronoun usage than did LEP students. The large variance component for students (nested within language background) in all three designs shows that there were substantial individual differences between students within a language group. Some students had mastered pronoun usage while others had not. The component for students, then, reflects the range of mastery of pronoun usage in the same.

Variation contributing to absolute error. Most of the variance components presented in Table 2 do not involve interactions with students or with language background. In Design I, the pronoun form facet accounts for the greatest variance (34.0%). Students found relative pronoun items to be very difficult. In fact, as can be seen in Table 3, LEP students performed at about chance level on all relative pronoun items except those measuring the nominative rule (with 5 response choices for each item, chance level is 20%).

Table 2 also shows a substantial effect for the context of the item, i.e., whether the sentence was embedded in a paragraph. The

Table 2
Proportion of Total Variation Accounted
For by Each Variance Component

Variance Component	Design I	Design II	Design III
Language Background [L]	12.3	19.6	19.4
Student [S(L)]	20.6	33.6	37.3
Pronoun Form [F]	34.0	[a]	—
Embedded [E]	4.1	26.2	17.3
Rule [R]	<3.5	<3.5	6.0
F E	5.1	—	—
F R	4.6	—	—
F S(L)	4.0	—	—
E S(L)	<3.5	6.6	<3.5
Residual	3.7	5.6	6.1
All others	11.6	8.4	13.9
Total	100.0	100.0	100.0

[a] Not applicable.

Note: Only variance components accounting for more than 3.5% of total variance are listed here.

variance component for embeddedness is smaller for Design I than for Designs II and III because the effect of relative pronouns (who-whom) overwhelmed that of embedding in Design I. The means in Table 3, 4, and 5 show that all students found it much more difficult to determine correct pronoun usage when the target sentence was embedded within other sentences. The difference in performance between embedded and non-embedded items was similar for FEP and LEP students.

Interestingly, the rule of grammar produced substantial variation in performance only in Design III. As Table 5 shows, students tended to perform worse on the items measuring the possessive rule than on items measuring the other rules. This effect appeared only when items measuring plural pronouns were included in the analysis (Design III), and not when singular items were included (Design II). As is indicated by the small variance component for rule in Design I (where the possessive rule was not included), student performance did not vary much across items measuring knowledge of the nominative and the three objective rules.

Table 2 shows two other effects in Design I that contributed to absolute error variance. The pronoun form facet interacted with the embeddness facet and with rules. The interaction between pronoun form and embeddedness indicates that the difference between performance on embedded and non-embedded items was not constant across relative and non-relative pronoun items. This result is clearly seen in Table 3. Both FEP and LEP students did much better on non-embedded items than on embedded items only when the pronouns were not in the relative

Table 3
Descriptive Results for Major Sources of Variation
in Design I

Factor	FEP[a]		LEP[a]	
	M	SD	M	SD
Non Relative Pronouns	.70	.15	.54	.18
Context				
Non-Embedded	.87	.17	.66	.24
Embedded	.53	.22	.42	.20
Rule				
Nominative	.59	.21	.49	.19
Direct Object	.66	.23	.51	.21
Indirect Object	.76	.23	.56	.28
Object of Preposition	.78	.18	.60	.29
Relative Pronouns	.34	.15	.24	.13
Context				
Non-embedded	.35	.17	.27	.14
Embedded	.33	.18	.22	.15
Rule				
Nominative	.57	.24	.52	.24
Direct Object	.20	.18	.13	.17
Indirect Object	.25	.21	.19	.21
Object of Preposition	.33	.28	.14	.20

[a] FEP = Fluent English Speaking. LEP = Limited English Speaking.

Table 4
Descriptive Results for Major Sources of
Variation in Design II

Factor	FEP		LEP	
	M	SD	M	SD
Context				
Non Embedded	.84	.17	.62	.24
Embedded	.51	.21	.38	.18

Table 5
Descriptive Results for Major Sources of Variation in Design III

Factor	FEP		LEP	
	M	SD	M	SD
Context				
Non-Embedded	.78	.17	.61	.23
Embedded	.56	.20	.42	.19
Rule				
Nominative	.62	.20	.49	.23
Direct Object	.68	.22	.53	.21
Indirect Object	.77	.23	.64	.25
Object of Preposition	.81	.20	.63	.31
Possessive	.46	.25	.27	.21

form. When the items called for relative pronouns, in contrast, student performance was very similar for non-embedded and embedded items. Thus, for relative pronoun items, the presence or absence of context did not affect student performance. Students performed poorly in both cases.

The interaction between pronoun form and rules can also be seen clearly in Table 3. For non-relative pronouns, students showed similar performance on all rules, with performance on nominative items somewhat lower than performance on objective items. For relative pronouns, on the other hand, performance on the nominative items was much higher than performance on the objective items. About half of the students knew when to use "who" (the relative form of nominative pronouns) but very few FEP students and no LEP students knew when to use "whom" (the relative form of objective pronouns). To determine whether students' performance differed across the three objective rules, Design I was also analyzed without the nominative rules (including only the three objective rules). The interaction between pronoun form and objective rules nearly disappeared (it accounted for only about 1% of the total variance), showing that students performed nearly the same on the three objective rules. Given this finding, then, it would not be necessary to retain information on the three objective rules. The mean for all objective items as an undifferentiated set would be sufficient.

Variation contributing to relative error. A notable feature of Table 2 is the lack of interactions between any facet and language background. This finding shows that the pattern of performance across the dimensions of the test among FEP students was the same as that for

LEP students. Coupled with the large component for language background, this result indicates that the profiles for the two groups have the same shape, with the profile for FEP students being higher than that for LEP students.

There were surprisingly few interactions between students and facets. The component for the interaction between students and pronoun form in Table 2 indicates that the rank order of students on relative items was not the same as the rank order of students on the non-relative items. There are two possible interpretations of this result. The first, which is highly unlikely given the huge main effect for pronoun form, is that some students found the relative pronoun item easier than the non-relative pronoun items while the rest found the non-relative pronoun items easier than the relative pronoun items. A far more likely interpretation is that the difference in performance between relative and non-relative pronouns was larger for some students than for others. It is unlikely that any students performed better on relative pronouns than on non-relative pronouns.

A similar interpretation can be given for the interaction between students and the embeddedness facet in Design II. Since it is unlikely that any student performed better on embedded items than on non-embedded items, the most likely interpretation of the interaction is that the difference in performance between embedded and non-embedded items was larger for some students than for others.

Finally, it should be noted that the residual variance component represents the interaction between all facets in the design, including students and language background, plus unsystematic error. A large

residual variance component usually reflects sources of variation that have not been taken into account in the measurement. The small magnitude of the residual component in all three designs in the present study suggests that all important test facets have been taken into account in the design of the test.

The Primary Sources of Variation and Example Diagnostic Profiles

The only sources of variation in test performance that exceeded 3.5% of the total variation were the pronoun form, embeddedness, and rule facets. The person (first vs. third) and the number (singular vs. plural) of the pronoun did not produce variation among students' test scores. That is, students showed equal mastery of first and third person pronouns and showed equal mastery of singular and plural pronouns. Furthermore, the effect for items was very small, indicating that students performed similarly on both items in each cell of the test design.

The findings protrayed in Table 2 and described above can be used to make recommendations about the optimal diagnostic profiles for pronoun usage for the sample in this illustrative study. Only the large effect contributing to relative error (those involving interactions with students) would need to be incorporated into the score profiles for individual students. Since only the who-whom and embeddedness facets interacted with students, the profile for individual students would need only to consist of the mean scores for relative pronoun items, non-relative pronoun items, embedded items, and non-embedded items. Example profiles for three randomly selected students appear in Figure 1.

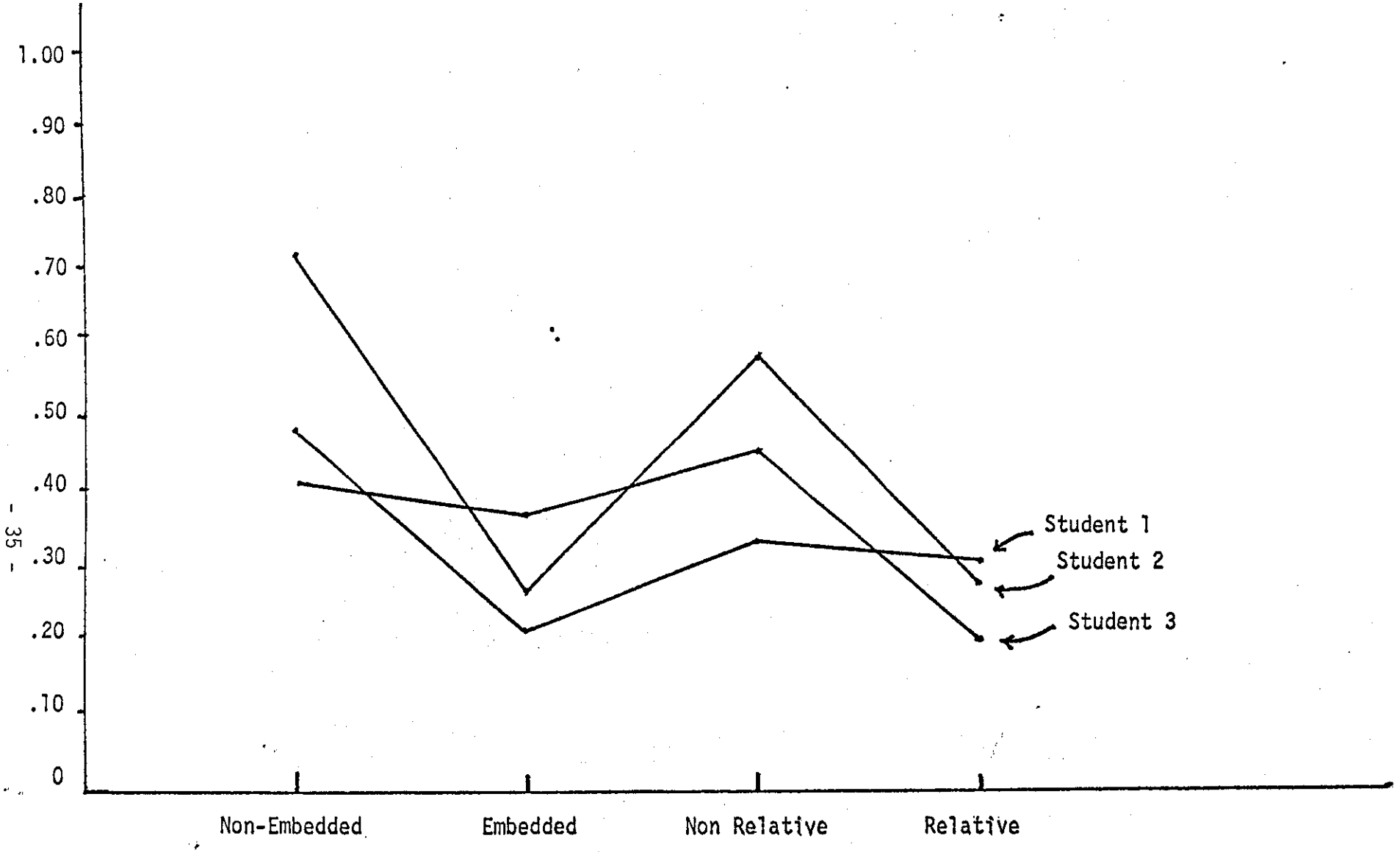

Figure 1. Individual Profiles for Three Students

The remaining large variance components--those contributing to absolute error but not relative error (components that do not involve interactions with students)--would guide the formation of class or group profiles. Since pronoun form interacted with embeddedness and rules, the group profile should present the means for embeddedness (embedded, non-embedded) and rules (nominative, objective, possessive) separately for relative items and non-relative items, as was done in Table 3. Figure 2 presents such profiles for FEP and LEP students. Since performance was similar across the three objective rules, only the mean score is presented for objective items. Furthermore, since performance was similar across person and number, first person plural, third person singular, and third person plural items were combined.

The mean profiles in Figure 2 show the general patterns of performance in this sample. Since the rule facet in the design did not interact with student, the means for nominative, objective, and possessive items are good representations of the performance of all students. This profile would show that all students need further instruction on the possessive rule and the objective rule for relative form. Similarly, the general pattern for embeddedness in Figure 2 suggests that students would need further instruction on all embedded pronouns and non-embedded relative pronouns.

In summary, the variance component analyses in the present study show that individual profiles for embedded and non-embedded and relative and non-relative pronouns, and a group profile for rules of grammar would be sufficient for diagnosing individual and group difficulties with pronoun usage. These profiles would be more

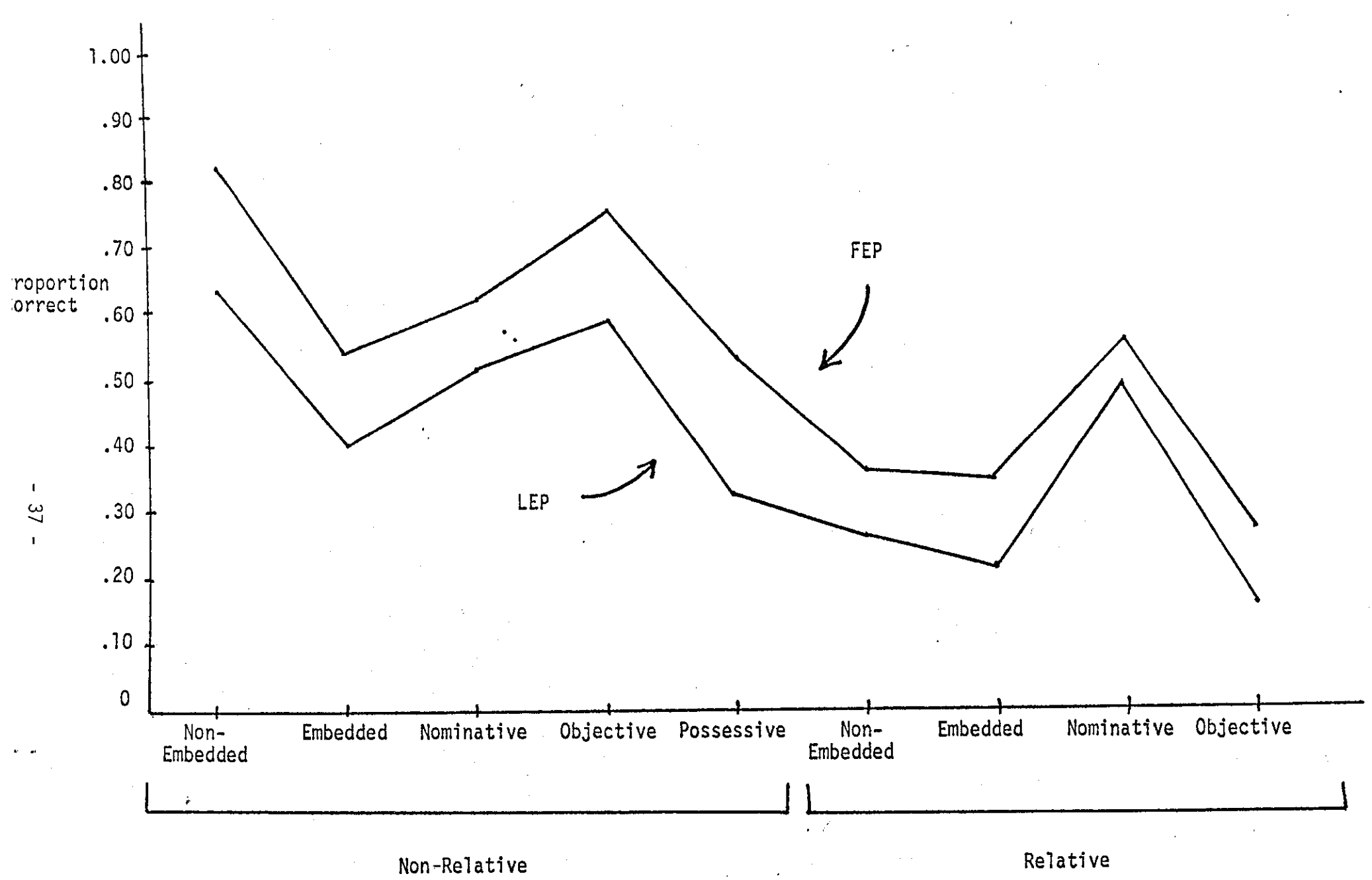

Figure 2. Mean Profiles for FEP and LEP Students

informative than the total score of the test, and suggest that diagnostic decisions based only on the total score might lead to erroneous consequences for the student and for the class. Not only do the variance component analyses show which aspects of test performance should be tabulated for individual and group diagnosis, they are also valuable for showing which aspects of pronoun usage do not need to be tabulated. Since student performance did not vary across number and person, these facets could be omitted from the diagnostic profiles.

The Optimal Number of Items

The previous section demonstrated how to use the relative magnitudes of the variance components to guide selection of scores for student and group profiles. This section reports the results of generalizability analyses to show how many items would have to be included in the test for dependable profile scores. The design of the generalizability analysis for each of the four scores in the individual profile (non relative, relative, non-embedded, embedded, see Figure 2) was students crossed with items. The items used in each generalizability analysis were all items in the original test that pertained to that score. The generalizability analysis of non-embedded items, for example, included all first-person, third-person, singular, plural, relative, non-relative, nominative, objective, and possessive items that were non-embedded.

The results of the generalizability analyses are presented in Table 6. Table 6 shows the number of items corresponding to different levels of generalizability. For example, it would take at least 10

TABLE 5

Number of Items Corresponding to Different Generalizability Coefficients

Score	Relative Decisions				Absolute Decisions			
	.50	.60	.70	.80	.50	.60	.70	.80
INDIVIDUAL PROFILE								
Non-Embedded	8	10+[a]	10+	10+	10+	10+	10+	10+
Embedded	9	10+	10+	10+	10	10+	10+	10+
Non-Relative	5	9	10+	10+	7	10	10+	10+
Relative	10	10+	10+	10+	10+	10+	10+	10+
GROUP PROFILE								
Non-Relative								
Non-Embedded	1	1	1	1	1	1	2	3
Embedded	1	2	3	7	6	9	10+	10+
Nominative	2	3	4	9	10+	10+	10+	10+
Objective	1	1	1	2	3	4	6	10
Possessive	1	1	1	1	2	2	4	6
Relative								
Non-Embedded	2	3	4	9	10+	10+	10+	10+
Embedded	1	2	2	4	4	7	10	10+
Nominative	1	1	1	1	2	2	3	5
Objective	1	2	2	5	2	2	4	8

[a] More than 10 items would be needed to obtain this level of generalizability.

items to measure individual proficiency on non-embedded pronouns with a .70 level of generalizability for relative and absolute decisions. On the other hand, only one item would be needed to measure group mean proficiency on non-embedded pronouns at the same level of generalizability for relative decisions (two items would be needed for absolute decisions).

Since the same items can be used to measure different aspects of pronoun usage, the total number of items in the test needed to obtain generalizable measures of each score in the profile is smaller than the sum of the number of items in Table 6. For example, an embedded, relative nominative item can be used to measure embedded pronoun usage, relative pronoun (who, whom) usage, and nominative pronoun usage. If the scores in the individual and group profiles were to be used for relative decisions (for example, selecting the bottom 20% of students for remedial instruction), a pronoun test of 20 items could be constructed so that each score in the profile had at least .70 generalizability. A pronoun test with the following configuration would satisfy the requirement listed in Table 6: 4 non-who-whom objective (embedded) items, 1 non who-whom possessive (embedded) item, 1 who-whom nominative (non-embedded) item, 2 who-whom objective (embedded) items, 3 other who-whom non-embedded items, 3 other embedded items (all who-whom) and 5 other non-embedded items (1 whom-whom and 4 non-who-whom). For absolute decisions, a pronoun test with 40 items could be constructed so that each score in the individual and group profiles had a level of generalizability of .70.

CONCLUSIONS

This paper described a four-step approach to constructing a diagnostic test that provides precise but practical information on students' problems and needs for additional instruction or remediation. The approach is based on analyzing the structure of the domain to determine which skills within the domain need to be assessed to diagnose students' problems.

The first step in the diagnostic process described here was to identify the factors that described the curricular domain (here, pronoun usage). Four content factors were identified: the rule of grammar (nominative, objective, possessive), the pronoun form (relative--who or whom, non-relative), the number (singular, plural), and the person (first, third). In addition, a factor corresponding to cognitive complexity was identifed: whether the context of the reading passage had to be taken into account to determine the correct pronoun. This factor was operationalized in two levels of embedding: a single sentence or a pragraph.

The second step was to construct a test with items representing all possible combinations of factors (content and cognitive complexity). Sensible items could be written for 46 combinations of factors. Two items were written for each combination, resulting in a 92 item test.

The third step in the diagnostic testing process used generalizability theory to determine which factors and interactions among them produced variation in students' scores. Specifically, the relative magnitudes of the variance components corresponding to all

factors and interaction among them in the test revealed which factors were important. This information was used to identify the information needed in diagnostic profiles. Only two content factors, rule and pronoun form, produced variation in student performance. The other two content factors, number and person, did not. Furthermore, cognitive complexity also had a large effect on student performance.

Some difficulties were common to all students (e.g., all students had more difficulty with possessive pronouns than with objective pronouns). This information could be entered in a single profile for the group or class. Other difficulties applied to some students but not others (e.g., some students did much worse on embedded items than on non-embedded items while other students performed similarily on both types of items). This information would be part of profiles for individual students. Since the number and person factors had no effect on student performance--all students performed about the same on singular and plural items on first-person and third-person items--there was no need to distinguish between these skills in the test or in the profiles.

Based on the information about the necessary ingredients of diagnostic profiles, the final step in the analytic process was to determine the minimum number of items needed to obtain a generalizable measure of each skill in the diagnostic profile. The results of the generalizability analyses showed that a 20-item test would be sufficient to measure mastery of pronoun usage if the teacher's

sufficient if the teacher's interest was in identifying each student's absolute level of mastery of each skill.

In short, the structure of the domain consisted of 46 skills in pronoun usage (all sensible combinations of the five factors). The inital test consisted of 92 items, 2 per skill. To adequately measure student performance on each of these 46 skills would probably take between 2 and 10 items per skill, resulting in an extremely long test. The analyses performed here showed that only 9 of the 46 skills need be assessed resulting in a vastly simplified and shorter diagnostic test.

Although the entire process of (1) identifying a domain, (2) constructing an initial test to fully represent the domain, (3) analyzing the performance on the inital test to determine the factors that influence student performance, and (4) constructing the final optimal test would be too time-consuming for a classroom teacher, the use of the final diagnostic test and score profiles would certainly be feasible for classroom practice. With a relatively short test (maximum of 20 minutes to administer, in this case), the teacher could identify students' strengths and weaknesses on all important aspects of the curriculum domain and make instructional decisions accordingly.

Specification of the domain structure underlying the test is an important issue in this diagnostic approach. It is important to specify the test as completely as possible. If factors in the test are left out, difficulties that students have on the test may be attributed to the wrong skills or may not be able to be identified at

all. Although complete specification is important, it is not necessarily difficult. In the present study, the generalizability analyses showed that only a small amount of variation in test performance was attributed to unexplained factors. Consequently, it is reasonable to conclude that all important factors in the domain were included.

Also important in domain specification is the need to consider aspects of the domain beyond content. Although several content factors did affect student performance, the cognitve complexity of the item had a major impact on performance. For example, even though many students could correctly identify when nominative pronouns should be used in a single sentence (a low level of cognitive complexity), many of them could not do so when the sentence was embedded in a paragraph requiring them to use the context of the paragraph (a high level of cognitive complexity). A teacher would come to different conclusions about mastery of pronoun usage from a test with items of low cognitive complexity compared to a test with items of high cognitive complexity. Without taking into account the influence of cognitive complexity on performance, the teacher may well make erroneous decisions about the need for additional instruction.

Finally, the results of the illustrative analyses presented here also have implications for taking into account multiple student populations. Teachers often give different tests to students from different population subgroups (for example, different language backgrounds), assuming that the performance of the groups is different. An implicit assumption, therefore, may be that some

groups excel on some material while other groups excel on other material; that is, that profiles of different groups may have different shapes. The strikingly parallel profiles of fluent English proficient students and limited English proficient students in the present illustrative study, however, raises a question about whether different tests are necessary. In this case, separate tests for each group would be unnecessary. To take into account the mean differences in performance between groups (fluent and limited English proficient students), the items measuring a particular skill on the diagnostic test could cover a range of difficulty (for example, varying the vocabulary level, or length of the sentences in a item).

REFERENCES

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

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

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

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

Chi, T.H., & Glaser, R. 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, 1980.

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

Doehring, D.G. & Aulls, M.W. The interactive nature of reading acquisition. Journal of Reading Behavior, 1979, 11(a), 27-40.

Edmonds, R. Making public schools effective. Social Policy, 1981, 12, 56-60.

Glaser, R. 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, 1970.

Hambleton, R. Item selection methods with criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

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

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

Kane, M.T., & Brennan, R.L. The generalizability of class means. Review of Educational Research, 1977 47, 167-292.

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

Millman, J. Computer-based item generation. In R.A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, Maryland: John Hopkins University Press, 1980.

Osburn, H.G. Item sampling for achievement testing. Educational and Psychology Measurement, 1968, 28, 95-104.

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

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

Quellmalz, E. Cognitive models for linking testing and evaluation. Los Angeles, California: Center for the Study of Evaluation, 1982.

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

Spiro, R.J. Constructive processes in prose comprehension and recall. In S.R.J. Spiro, B.C. Bruce & W.F. Brewer (eds.), Theoretical issues in reading comprehension perpectives from cognitive psychology, linguistics, artificial intelligence and Education. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1980.

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

Webb, N.M., & Shavelson, R.J. Generalizability of general educational development ratings of jobs in the U.S. Journal of Applied Psychology, 1981 66, 186-191.