DOCUMENT RESUME

ED 252 556 TM 850 025

AUTHOR Webb, Noreen; Herman, Joan
TITLE Diagnosing Students' Errors from Their Response Selections in Language Arts. Methodology Project.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 84
GRANT NIE-G-84-0112-P1
NOTE 24p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adaptive Testing; Computer Assisted Testing; *Diagnostic Tests; *Error Patterns; Grade 6; Intermediate Grades; *Item Analysis; *Language Arts; *Latent Trait Theory; Pronouns; *Test Construction; Testing Problems; Test Theory

ABSTRACT

This paper describes the development of a language arts test to assess the consistency of student response patterns and the feasibility of using the test to diagnose students' misconceptions. The studies were part of a project to develop computerized adaptive testing for the language arts with software to diagnose student errors. The domain-referenced test on pronoun usage featured matched pairs of test items. To diagnose errors it was assumed that (1) a misconception could only lead to an incorrect response; and (2) matched items would get the same incorrect response. The test was administered to three samples of inner-city sixth graders, in groups, and individually with students asked to explain the reasons for their responses. Results indicated that neither students' wrong answer choices nor the rationale for their answers were consistent enough to warrant analyzing student response patterns in the classroom. Student inconsistency in test answering strategies may have important implications for measurement models and theories that assume a systematic and analytic approach to test taking. (BS)

DELIVERABLE - NOVEMBER 1984

METHODOLOGY PROJECT

Diagnosing Students' Errors From Their Response Selections In Language Arts

Noreen Webb and Joan Herman

Study Directors

Grant Number

NIE-G-84-0112, P1

Center for the Study of Evaluation

Graduate School of Education

University of California, Los Angeles

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

Diagnosing Students' Errors From Their Response Selections

In Language Arts

The present set of studies were part of a project designed to develop a computerized adaptive test with a built-in computer program to diagnose errors in a subject matter area not explored in this way before: language arts. Specifically, the intent of the test was to uncover students' erroneous rules of usage to guide instruction and remediation in the classroom. As a first step in this process, a language arts test was developed and modified to assess the consistency of student response patterns and, consequently, the feasibility of using the test for diagnosing students' misconceptions.

The work of Brown and Burton and colleagues and Tatsuoka and colleagues shows that analysis of response patterns can be used successfully to diagnose students' errors in arithmetic. Brown and Burton (1978; see also Brown and VanLehn, 1980) use the answers to multiple items to determine the bugs (erroneous rules) that students exhibit in solving arithmetic problems. Their work recognizes that students can have misconceptions and still give correct answers to some or even most items. Furthermore, different misconceptions will lead to different numbers of right and wrong answers and also to different pattern specific responses. From the specific patterns of responses, Brown and Burton can determine a student's particular bug or bugs. The drawbacks to their approach at present are that the process is very complex, requires fairly large number of items, and has been applied to quite restricted domains, namely addition and subtraction of whole numbers.

The work of Tatsuoka and Birenbaum (Birenbaum & Tatsuoka, 1982,

1983; Tatsuoka, 1983; Tatsuoka, 1983) is similar to that of Brown and Burton in that the pattern of responses is of central concern. With modified scoring of arithmetic items, specifically taking into account the absolute value and the sign in the addition and subtraction of signed numbers, Tatsuoka (1983) uses item response theory to show how to calculate the likelihood oeach bug. As in Brown and Burton's work, erroneous rules can lead to different patterns of right and wrong responses across items, making the process of diagnosis complex.

The present set of studies tried to expand diagnosis of erroneous rules to a new subject matter domain, language arts, and to design a test with a simple way of diagnosing misconceptions, one that could easily be used in theclassroom. The domain is not only a novel one, but is broader than the domains of addition and subtraction used previously. Specifically, a test was designed to diagnose erroneous rules of pronoun usage. The test featured two major simplifications in its approach to diagnosing errors: (1) a misconception could lead only to an incorrect response, and (2) for parallel or matched items, a misconception would lead to the same incorrect response. Most importantly, the choice of response would immediately point to the misconception a student held. This approach to diagnostic testing could easily be adapted to a computerized setting giving the teacher a listing of the misconceptions for each student or for groups of students. This information could then be used to guide instruction and remediation.

The approach outlined above depends on consistency in student behavior namely that students do hold systematic misconceptions which they apply consistently to items of the same kind. If students do not consistently make the same kind of error, then it would be difficult, if

not impossible to identify their misconceptions. This paper describes the design and administration of a domain-referenced test in language arts and examines the consistency of student responses to test items. In addition to examining the response choices, the study also used introspective recall to investigate the consistency of students' reasoning.

METHOD

Development of the Test

To select the content of the test, language arts teachers were asked to indicate the kinds of grammar problems that students exhibited most frequently at the upper elementary grade levels and the areas in which a diagnostic test would be most helpful. The most common response was pronoun usage. The original diagnostic test developed in this project, then, measured pronoun usage. Discussions with language arts experts and examinations of language curricula and textbooks showed that four content factors represented an important domain of pronoun usage: pronoun rule (nominative, three types of objective, possessive), pronoun form (relative--who or whom, non-relative), pronoun number (singular, plural), and pronoun person (first, third). A fifth factor represented the cognitive complexity of the item, here based on whether students had to use the context of a reading passage to determine the correct pronoun. For one level of cognitive complexity, the pronoun was embedded in a single sentence and the referent was given. For the other, the pronoun was embedded in a short paragraph and students had to use the context to identify the referent.

The original test developed to represent the entire domain of pronoun usage had 92 items: 2 parallel (or matched) items for each of 46

combinations of the five factors. The matched items in a pair had identical grammatical structures, comparable vocabulary, and often had similar content.

Matched items also had the same distractors, representing common student errors related to rule, form, number, and person. A few exceptions were pairs of items in which the referent for the pronoun was male in one item and was female in other item. (Students did not have any difficulty inferring the correct gender, so the fact that one item called for a male pronoun and the second item called for a female pronoun did not influence student performance.) For example, if the first item in a matched pair called for the pronoun "him", one of the distractors would be "he"; similarly, the correct answer for the other item would be "her" and one of the distractors would be "she". Distractors with different genders appeared only once in a matched pair of items; all other distractors for both items were identical. For the majority of matched items on the test (62%), all five responses were identical.

By design, then, students who held erroneous rules of pronoun usage were expected to give the same response (or comparable responses) to both items. An example pair of matched items is the following:

(1) David, _____ the coach had given another chance, worked very hard all summer.

(2) The toddler, _____ the grandfather had given cake, was covered from head to toe with frosting.

In the above example, a student who responded "who" (incorrect pronoun rule: nominative instead of objective, a common error) was expected to give the same incorrect response to the second item. This response would indicate incorrect pronoun rule, but would indicate correct pronoun form (relative). Similarly, students giving the answer "he" to the first item were expected to give "he" (or "she"; a toddler could be female) for the

second item, but correct pronoun rule (nominative).

The following example involves pronouns embedded in a paragraph.

(3) Mr. and Mrs. Roberts were on their way to the airport when Mr. Roberts realized that he had left his calculator at home. Fortunately, Mrs. Roberts loaned _____ her calculator.

(4) Julia and Sandy were on their way to the game when Sandy realized that she had left her pompoms at home. Fortunately, the drill coach loaned _____ some extras.

In this example, students making an error in the first item were expected to make the same error in the second item. Students giving the answer "he" in the first item (incorrect pronoun rule: nominative instead of objective) were expected to give the answer "she" to the second item. (Students did not have any difficulty inferring the correct gender, so the fact that the first item called for a male pronoun and the second item called for a female pronoun did not influence student performance.) In this way, it would be relatively straightforward to identify students' misconceptions about pronoun usage. In the latter example, giving the answers "he" and "she" would indicate incorrect pronoun rule (nominative vs. objective), but would indicate correct person (third), correct form of the pronoun (non-relative), and the ability to determine the correct referent for a pronoun embedded in a paragraph. Similarly, giving the answer "them" to both items (which a few students did, thinking that the pronoun referent was the first two people named in the paragraph) would indicate difficulty determining the correct referent for pronouns embedded in a paragraph, but would indicate correct form (non-relative). These inferences about students' misconceptions could then be used to guide instruction and remediation.

Analysis of student performance on the original test

(generalizability analysis--see Cronbach, Gleser, Nanda, & Rajartnam, 1972--and analysis of variance) showed that students performed similarly on singular and plural pronouns and on first and third person pronouns, but performed differently on nominative and objective pronouns, on relative and non-relative pronouns, and on embedded and non-embedded pronouns. (For a complete description of these analyses and results, see Webb, Herman, & Cabello, 1983.) Consequently, a subset of 16 items was selected from the original test to represent all combinations of nominative and objective pronouns, relative and non-relative pronouns, and embedded and non-embedded pronouns. For each of the 8 combinations of these pronoun variations, there were 2 matched items, as described above. All pronouns were singular (rather than plural) and in the third person (rather than first). This 16 item test is the focus of the current set of studies.

Test Administration

Sample. The test was administered to three samples of sixth-grade students (n=79, n=26, n=21) from schools in an inner-city school district The schools are located in a low- to middle-SES area with a high rate of transiency and a mixed population (90% Hispanic, 6% Black, 2% Asian, and 2% non-minority whites). All students were classified by the district as Fluent English Proficient.

Procedure. Test administration varied for the three samples. For the 79-student sample (hereafter called Study 1), the test was administered conventionally in groups. Each item had five alternatives: one correct and four distractors. Students were permitted to ask the test administrator about difficult vocabulary, but were not permitted to ask other questions or consult with anycne else.

For the other two samples of students, the test was administered

individually. After each item, students were asked to explain why they selected their response (introspective recall). Rationales given included indicating the correct referent of the pronoun, or characteristics of the correct referent such as gender or number. Sometimes students could not give a reason or could only give a vague reason such as "It sounds good". These responses were not classified as rationales for the analyses.

The only difference between the two latter administrations was the response format. For the 21-student sample (Study 2), the same multiple choice format was used as in Study 1. For the 26-student sample (Study 3) a multiple-choice format was not used. Instead, students were asked to construct their response choices. The purpose of the varied formats was to determine whether providing students with alternative responses influenced the consistency of their response selections or the consistency of their rationales for selecting responses.

RESULTS

Consistency of Students' Response Selections

The data on consistency of students' response selections for Studies 1, 2, and 3 are presented in Tables 1, 2, and 3. The numbers without parentheses are the proportions of students in each sample who gave each combination of correct and incorrect responses in a matched pair of items: both items correct, both incorrect with the same error, both incorrect but with different errors, and one item correct and the other incorrect.

Prior to examining the data in each table, statistical tests were performed to determine whether the performance of the three samples were comparable. The statistical test used here was the chi-square test comparing proportions across independent samples (see Fleiss, 1981,

p. 139). This test was computed for each of the 36 proportions in the tables. Due to the large number of statistical tests being conducted (36) a significance level of .01 was used instead of the conventional level of .05. Only two of the 36 tests were significant. For these two comparisons, post hoc analyses were conducted to determine the precise differences between studies (see Fleiss, 1981). First, a higher proportion of students gave the correct answer to both non-relative, embedded, objective items in Study 3 than in Studies 1 and 2. Second, a higher proportion of students gave the correct answer to both relative, embedded, nominative items in Studies 2 and 3 than in Study 1. The small number of significant comparisons, coupled with the inconsistent directions of the comparisons, suggests that the three studies are comparable. It was concluded that the method of test administration and the format of the items had little effect on student performance.

As can be seen in Tables 1 through 3, students were rarely consistent in their responses to each pair of matched items. If students were consistent in their conceptions about pronoun usage, they would have given the same or comparable responses (correct or incorrect) to both items in a matched pair. The number of students giving different incorrect answers to the items or giving the correct answer for one item and the incorrect answer for the other item would have been small. In all three studies, however, the proportions of students giving consistent responses were substantial. In Study 1, 44% of the students, averaging over all pairs of matched items, were inconsistent in their responses to matched items (15% making different errors + 26% giving one correct response and one incorrect response). The average percentages of inconsistent responses for Studies 1, 2, and 3 were 37% (5% + 32%) and 28% (11% + 17%), respectively. These data suggest that substantial

numbers of students did not hold systematic misconceptions of pronoun usage.

A few examples of inconsistent responses across matched items are instructive to show that such performance is unlikely to occur with systematic misconceptions about pronoun usage. First, consider the second pair of items given in the section of this paper entitled "Development of the Test." The most common instance of one correct and one incorrect response was "them" for the first item (incorrect referent) and "her" for the second item (correct). Second, the most common combination of different incorrect responses was "them" for the first item (incorrect pronoun referent and incorrect rule: nominative vs. objective). It is difficult to conceive of systematic erroneous rules for pronoun usage that would produce these combinations of responses. For embedded items, rules for selecting the response, other than misconceptions about the type of pronoun or pronoun referent, were considered. An obvious example is a student selecting the first response alternative (for the multiple choice format) that would be grammatically correct if the sentence with the pronoun were not embedded in a paragraph. While there were a few isolated instances of this behavior, it could not explain most student behavior.

By far, the most important information for diagnosing misconceptions is the proportion of students making the same error on both items in a matched pair. Students making the same error on matched items very likely are using an erroneous rule consistently to determine the correct pronoun. From the error made, it is easy to identify the erroneous rule. The teacher can use this information accordingly. The proportions of students making the same errors on matched items were low, however, on the average, across studies, ranging from 9% for Study 2 to 18% for Study

3. Even among individual pairs of matched items, the proportions were rarely large enough to be of practical value in the classroom. One of the few large proportions, for example, was the 54% of students in Study 3 who gave the same incorrect answer to the two items measuring knowledge of objective, relative pronouns embedded in a paragraph. All of these students gave the answer "who" instead of "whom", showing that they had a misconception concerning the nominative and objective cases of relative pronouns. Large proportions occurred too infrequently, however, to warrant such detailed testing. Overall, then, the data in Tables 1 through 3 show that the analysis of incorrect responses would be beneficial only for a small number of students and only for a few items on this test. Analyzing incorrect responses probably has little value for the classroom, at least in pronoun usage.

Consistency of students' rationales for selecting responses. It was hoped that students who made consistent response choice (both correct or the same incorrect choice both times) would be more consistent in the rationales for their selections than would students who were inconsistent in their responses (different incorrect responses or one correct and one incorrect response). Such results would make it reasonable to use students' rationales to make inferences about their particular misconceptions.

The data on consistency of students' rationales appear in Tables 2 and 3. The numbers in parentheses are the proportions of students who gave each response combination who also gave the same rationale for giving each response. For example, the first entry in Table 2 shows that of the 95% students who gave the correct answer to both items in this pair, 15% of them gave the same rationale for both items. In terms of

numbers of students, 95% of 21 students is 20 students; and 15% of 10 students is 3 students. Similarly, for the third entry in the first column, of the 29% of students who gave the correct answer to both items in this pair, 17% of them were consistent in the rationales for selecting their response. In terms of numbers of student, 29% of 21 students is 6 students; and 17% of 6 students is 1 student. So, the percentage of students giving consistent rationales often corresponds to very small number of students in Tables 2 and 3.

As was done for the consistency of students' response selections, statistical tests were performed to determine whether the students in Studies 2 and 3 differed in their consistency of rationales. None of the statistical tests was statistically significant, showing that the two samples were comparable.

As the data in Tables 2 and 3 show, students who gave consistent answer choices did not give more consistent rationales than students who did not give consistent answer choices. Averaging over the two studies on over all pairs of matched items, only 33% of the students who gave consistent answer choices gave the same rationale for both items in a pair. By comparison, 27% of the students who gave inconsistent answer choices across matched items gave the same reasons for making their choices. The difference between percentages is not statistically or practically significant.

In summary, neither students' wrong answer choices nor their rationales for giving their answers were consistent enough in the present study to warrant analyzing patterns of students' responses in the classroom. Such analyses would produce little information that would help teachers diagnose students' misconceptions about pronoun usage.

CONCLUSIONS

There are several possible explanations for the lack of consistent student responses in the present set of studies. First, students may have held systematic misconceptions about pronoun use but behaved randomly or carelessly on the test. If this hypothesis were true, the high degree of inconsistency would suggest that students were frequently careless or random in their responses on many items. The procedures of the two studies using introspective recall, however, tend to refute this hypothesis. The one-to-one testing situation most likely facilitated attention, and the fact that students were required to give a rationale for their responses should have encouraged care and thoughtfulness. Students often gave rationales for selecting their responses, even if their rationales were not consistent across items.

Second, the items may not have been sufficiently parallel in structure to elicit consistent student behavior. That is, features of the items, such as vocabulary, content, and grammatical structure, may have led students to use different strategies for selecting pronouns in different items. Given the careful control of vocabulary, content, and grammatical structure in the matched items, however, this hypothesis also seems unlikely.

One qualification of the hypothesis just mentioned concerns the grammatical structure of the items. While items were carefully matched on surface structure, it was difficult to ensure that items were matched on underlying or deep structure. Slight differences in surface structure may have signalled different deep structures among matched items, prompting students to give different rationales for their responses (see Paivio & Begg, 1981). For example, items 3 and 4 (given in the section of this paper describing the development of the test) had two slight

differences in surface structure which may have influenced the way students processed the item. First item 3 had two noun agents ("Mr. and Mrs. Roberts"), whereas item 4 had three ("Sandy, Julia, and the drill coach"). The additional noun agent in item 4 may have made that item more difficult to process (see Schank, 1982). Second, in item 3, the pronoun to be identified was followed by a pronomial phrase ("her calculator"), whereas in item 4, the pronoun to be identified was followed by an adverbial phrase ("some extras"). The adverbial phrase in item 4 required the student to take an extra step to process the item, namely determining that the "extras" referred to the "pom poms". Taken together, these differences in surface structure may have made item 4 more complex than item 3. While it seems reasonable that students with specific rules of pronoun usage would not be affected by such differences in grammatical structure, this hypothesis has never been examined empirically and so cannot be ruled out a priori.

Third, a greater number of items may be needed to obtain consistent responses. In the test developed here, there were only two matched items per topic of pronoun usage. One could argue that students would have demonstrated greater consistency across more items. The careful matching of the items and the lack of consistent reasons given by students for selecting responses argue against this hypothesis. However, further studies should use more items per topic in the domain to test this hypothesis.

The fourth hypothesis, and the most persuasive, is that students' misconceptions about pronoun usage were not well defined or precise. That is, students may not have had any well-articulated rules upon which to draw, and thus could not apply any rules consistently. The

inconsistency of students' rationales for selecting their responses tends to support this hypothesis. For matched items, students often gave one rationale for selecting the pronoun in one item (such as identifying the gender of the referent) and using a different rationale, a vague one, or no rationale for the other item (such as identifying the number of the referent, merely saying "It sounds good", or saying "I don't know"). Sometimes students identified the correct referent in one item and identified the wrong referent in the other matched item. For items so carefully matched in grammatical structure, this behavior suggests strongly that students did not consistently use systematic misconceptions or erroneous rules to answer the test items.

If students do not use consistent strategies for answering test items in pronoun usage, probably amongst the most well-ordered subject area of language arts, then analyzing patterns of students' responses to identify errors may not be possible in language arts. What remains to be tested is whether analysis of students' response patterns can be successfully applied in other areas outside of mathematics.

The results of the present set of studies also have implications for other measurement issues. The first relates to the diagnostic strength of item distractors (Roid and Haladyna, 1982) and their use in adaptive testing, where decisions about branching depend on the specific answer a student gives for an item or set of items (Roid, 1969; Swinton, 1984). In such adaptive testing situations, decisions for branching made on the basis of specific responses depend on placing confidence in the student's response. If students' responses do not correspond to systematic misconceptions, as the data in this paper suggest for language arts, such branching decisions may often be misleading or erroneous.

A second implication is related to assumptions about students' test-taking behavior which apparently underlie certain measurement models and theories. The feasibility and usefulness of answer-until-correct (Wilcox, 1983) and confidence marking models (LeClerq 1981), for example, seem to assume a systematic and analytic approach to test tasks, where students are able to rule out, or probabilize, their responses on the basis of systematic analysis of both problems and available responses. The results of this study suggest that such assumptions may be untenable for young students. An analytic approach would imply that students go through a parallel set of steps in answering each item, scanning and categorizing each problem in terms of a consistent set of key features, using a consistent cognitive model. The rationales which students provided, however, indicated that they did not attend to the same key features in responding to parallel items: stimuli which were salient in one problem were not in the next and thus students did not respond in a uniform fashion. Their responses were not random, but their conceptualization of the problem seemed to vary, perhaps as a function of carelessness, perhaps as a function of an incomplete or inconsistent cognitive model. Further, students seemed not to attend to item distractors to help them formulate their response, i.e., there was no evidence that students arrived at an answer by ruling out certain alternatives.

While these results may raise important questions about the comprehensiveness and usefulness of some new theories applied to younger children, the problem may lie with the mental model underlying the test. One might argue, for example, that students' behavior appeared random because it was evaluated against an inappropriate mental model and/or that the model was at too gross a level of generality. Even though the test domain was very carefully developed, based on curricular and

instructional structure and embodying teachers' perceptions of the types of and reasons for student errors, such a criticism cannot be overruled. However, given the care exercised in test development and the attention to the available research base, one must question whether it is possible and/or feasible, given the current state of the research, to build classroom tests in a variety of subject areas which reflect appropriate mental models.

The studies reported here investigated the development of a diagnostic test which would provide information not only about students' attainment of particular skills and/or objectives but would also help to identify the sources of their errors and thus provide teachers with concrete guidance in designing remedial strategies. The findings lead to pessimism regarding the feasibility of such an approach in routine classroom practice, in terms of the resources required for test development, the level of precision likely to be attained given the current state of the art, and the number of items and student testing time which would be required. On a more specific level, the results are discouraging as to the current feasibility and utility of deriving information beyond traditional right-wrong scoring from student test responses in areas other than mathematics.

REFERENCES

Birenbaum, M. & Tatsuoka, K. K. (1982) On the dimensionality of achievement test data. Journal of Educational Measurement, 19, 259-266.

Birenbaum, M. & Tatsuoka, K. K. (1983) The effect of a scoring system based on the algorithm underlying the students' response patterns on the dimensionality of achievement test data of the problem solving type. Journal of Educational Measurement, 20, 17-26.

Brown, J. S. & Burton, R. R. (1978) Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 2, 155-192.

Brown, J. S. & Burton, R. R. (1980) Repair theory: A generative theory of bugs in procedural skills. Cognitive Science, 4, 379-426.

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

Fleiss, J. L. (1981) Statistical Methods for Rates and Proportions. (Second Edition) New York: Wiley.

LeClerq, D. Confidence Marking. Monograph of the Laboratoire de Pedagogie Experimentale, Universite de Liege (Belgium), 1981.

Paivio, A. and Begg, J. (1981) The Psychology of Language. New York: Prentice Hall.

Roid, G. (1969) Branching methods for constructing psychological test scales. Unpublished doctoral dissertation, University of Oregon.

Roid, G. and Haladyna, T. (1982) A technology for test item writing, New York: Academic Press.

Schank , R. C. (1982) Reading and understanding: Teaching from the perspective of artificial intelligence. Hillsdale, NJ: Lawrence Erlbaum Association.

Swinton, S. S. (1984) The art of not asking questions. Unpublished paper.

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

Webb, N. M., Herman, J., & Cabello, B. (1983) Item Structures for Diagnostic Testing. Center for the Study of Evaluation, UCLA Graduate School of Education.

Wilcox, R. (1982) Some empirical and theoretical results on answer-until-correct scoring procedures. British Journal of Mathematical and Statistical Psychology.

Table 1

Percentage of Students in Study 1 with Each Response Pattern

Matched Pair of Items	Both Correct	Both Incorrect		One Correct, One Incorrect
		Same Error	Different Error	
Non-Relative				
Non-Embedded				
Nominative	90	5	2	2
Objective	85	1	1	13
Embedded				
Nominative	14	27	23	37
Objective	39	6	15	39
Relative				
Non-Embedded				
Nominative	49	6	6	38
Objective	8	35	27	30
Embedded				
Nominative	32	11	16	40
Objective	9	29	27	35
Average over all Measures	41	15	15	29

Note: $n = 79$

Table 2

Percentage of students in study 2 with each response pattern

Matched Pair of Items	Both Correct	Both Incorrect		One Correct, One Incorrect
		Same Error	Different Error	
Non-Relative				
Non-Embedded				
Nominative	95(15)[a]	0(-)[b]	0(-)	5(0)
Objective	95(10)	0(-)	0(-)	5(0)
Embedded				
Nominative	29(17)	19(50)	0(-)	52(18)
Objective	57(33)	0(-)	5(0)	38(0)
Relative				
Non-Embedded				
Nominative	52(27)	0(-)	10(50)	38(0)
Objective	14(33)	29(33)	10(0)	48(20)
Embedded				
Nominative	67(36)	5(0)	5(0)	24(20)
Objective	29(17)	19(25)	10(50)	43(11)
Average over all Measures	55(24)	9(27)	5(20)	32(9)

Note: $n = 79$
a: percentage of the students in this cell who gave the same rationale for choosing both answers.
b: because no students gave this combination of answers, it was not possible to examine consistency of rationales.

Table 3

Percentage of students in study 3 with each response pattern

Matched Pair of Items	Both Correct	Both Incorrect		One Correct, One Incorrect
		Same Error	Different Error	
Non-Relative				
Non-Embedded				
Nominative	100(42)[a]	0(-)[b]	0(-)	0(-)
Objective	88(30)	0(-)	0(-)	12(33)
Embedded				
Nominative	35(56)	27(14)	12(0)	27(57)
Objective	80(30)	4(100)	4(100)	12(0)
Relative				
Non-Embedded				
Nominative	62(53)	8(100)	0(-)	29(28)
Objective	5(100)	43(22)	33(29)	19(75)
Embedded				
Nominative	64(50)	4(0)	4(0)	27(17)
Objective	4(0)	54(54)	33(12)	8(0)
Average over all Measures	55(45)	18(48)	11(28)	17(30)

Note: $\underline{n} = 79$

a: percentage of the students in this cell who gave the same rationale for choosing both answers.

b: because no students gave this combination of answers, it was not possible to examine consistency of rationales.