Assessing Student Representations of Inferential Statistics Problems

CSE Technical Report 553

Nancy C. Lavigne and Robert Glaser

CRESST/Learning Research and Development Center
University of Pittsburgh

December 2001

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.1. Models-Based Assessment Design: Individual and Group Problem Solving—Theory Into Practice. Robert Glaser, Learning Research and Development Center, University of Pittsburgh / CRESST, and Gail Baxter, Educational Testing Service, Project Directors.

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

ASSESSING STUDENT REPRESENTATIONS OF INFERENTIAL STATISTICS PROBLEMS

Nancy C. Lavigne and Robert Glaser[1]
CRESST/Learning Research and Development Center,
University of Pittsburgh

Abstract

A student's understanding of a problem before he or she solves it is a critical component of successful problem solving. This understanding is based on how a problem is represented; that is, whether a problem is understood in terms of principles or solution methods or whether the focus is on features that are irrelevant to its solution. In this study, we examined 12th-grade students' representations of statistics problems for which they used a sorting task. We indexed the degree to which problems were sorted according to statistical methods, and we used a verbal protocol to assess students' explanations of their sorts. Two main findings are discussed: (a) Students' problem representations were based on key features that underlie statistical methods, and the extent to which these features were stressed in explanations varied; and (b) students' explanations were more sensitive to the nature of their problem representations than was their sorting performance. These findings have implications for the design of instruction and assessment.

The objective of this study was to examine a significant aspect of problem solving, namely, problem representation in the domain of statistics. Problem solving is a critical component of the mathematics curriculum (National Council of Teachers of Mathematics [NCTM], 1989; 2000) and has been the focus of research for two decades. Yet, comprehensive problem-solving programs are not always implemented in K-12 classrooms, and research still is needed to fully understand the complexities of the problem-solving process (Lester, 1994). It is a well-established finding that solving a problem successfully requires that a solver first understand the problem, and second, perform the appropriate procedure (Brenner et al., 1997; Mayer, 1985; Mayer, Lewis, & Hegarty, 1992). Problem understanding involves having an accurate *problem representation*, which results from a process of trying to connect one's content knowledge to requirements of the specific problem under consideration before executing a solution procedure (Silver & Marshall, 1990).

[1] We gratefully acknowledge Joy Evans' assistance with the data collection and thank Corinne Zimmerman, Marsha Lovett, Marguerite Roy, Gloria Berdugo, and Maria Magone for helpful comments on an earlier version of this draft.

Inability to represent a problem accurately can result in unsuccessful problem solving (Marshall, 1995; Mayer et al., 1992). This difficulty is particularly common to word problems because their solution depends on a problem representation that is constructed from verbal and contextual information provided in the text. The ability to represent problems accurately by translating words into appropriate equations and interpreting the solutions into problem contexts is critical to understanding and solving problems, particularly when they are complex (Koedinger & Nathan, 1999). One way to facilitate the development of problem representation skills is to give students experience with word problems before they learn how to manipulate symbols (Brenner et al., 1997; Koedinger & Nathan, 1999; Nathan & Koedinger, 2000). Students can thus acquire verbal problem representation skills, which can provide the basis for representing the constraints of problems in symbolic language, before actually solving equations (Koedinger & Nathan, 1999). In other words, once acquired, students' verbal problem representations can scaffold the development of symbolic representations prior to problem solving.

Research examining how individuals represent word problems has centered on chess, physics, and mathematics. These investigations have revealed the relationship between accuracy of problem representation and expertise (Chase & Simon, 1973; Chi, Feltovich, & Glaser, 1981; Larkin, McDermott, Simon, & Simon, 1980; Schoenfeld & Herrmann, 1982; Silver, 1981). An important aspect of this relationship is the ability to perceive the structure of a subject matter domain; that is, to understand the concepts and how they are interrelated. The expertise literature reveals that novices (or students in the early phases of learning a particular content) do not perceive the domain structure in problems. Rather, they focus on irrelevant features of the problem, such as the story line or content not pertinent to solving the problem. In other words, novices represent problems in terms of surface features. In contrast, experts (or students with extensive experience in a particular content domain) represent problems in terms of relevant principles or methods—the structural features—and apply them appropriately.

The notion that individuals represent problems in terms of surface or structural features is consistent with schema theory. According to this theory, a problem is represented based on knowledge of a problem type (Chi et al., 1981). A schema results from experience in solving problems that share common features (Marshall, 1995). The most relevant features are abstracted and incorporated into existing schemata or form the basis of new schemata. Successful problem solving thus requires that critical problem features be recognized and mapped onto existing schemata (Marshall, 1995; Mayer et al., 1992). A common methodology for examining problem representation is problem categorization on a sorting task (Chi et al., 1981; Mariné & Enscribe, 1994; Quilici & Mayer, 1996; Schoenfeld & Herrmann, 1982; Silver, 1981). Information about students' problem representations can be gleaned from their ability to categorize problems according to basic problem types (Mayer et al., 1992).

Although a substantial amount is known about students' problem representations in physics and mathematics, little work has been done in the domain of statistics. The cognitive research in this area is young, and much has yet to be learned about how students solve statistics problems (Becker, 1996). Statistics is a particularly fruitful domain of investigation given its complexity and reputation for being hard to grasp (Cumming, Thomason, & Zangari, 1995; Garfield & Ahlgren, 1988; Shaughnessy, 1996), its importance in everyday decision making (Hauff & Fogarty, 1996; Moore, 1997), its current role in the K-12 mathematics curriculum and emphasis on problem solving, and the instructional dependence on word problems. There is a need for studies that examine students' ability to accurately represent statistics problems.

Students' ability to structure their knowledge of a domain often can depend on how the curriculum is organized. The manner in which statistics is treated in the mathematics curriculum can vary across schools. Statistical content can be embedded within the mathematics curriculum at various points in the year or treated as a separate course. The key in each of these contexts is whether the relationship between concepts is emphasized and when. Content related to hypothesis testing, for instance, often is taught as a course, and the curriculum unit is organized in such a way that principles are taught first (e.g., central limit theorem and probability), followed by hypothesis testing methods. Each method is taught in isolation from the others (Lovett, 2001). Hypothesis testing usually is taught at the university; however, some high schools have provided students with opportunities to learn about inferential statistics (e.g., Advanced Placement statistics courses). As in mathematics, statistics instruction traditionally has focused on developing students' ability to solve equations. It rarely allows students to practice making decisions about appropriate analyses (Lovett, 2001; Lovett & Greenhouse, 2000). The common and distinguishing structural features that underlie various hypothesis tests are therefore not addressed explicitly. Consequently, it is often difficult for students to represent problems in terms of the appropriate structure (Hubbard, 1997). The manner in which the curriculum is organized can thus make it more difficult for students to perceive that structure in the problems they are asked to solve.

Frequent use of word problems and increasing emphasis on problem- and project-based methods in statistics classrooms (e.g., Derry, Levin, Osana, & Jones, 1998; Fillebrown, 1994; Lajoie, Lavigne, Munsie, & Wilkie, 1998; Lavigne & Lajoie, 2000) enhance the value of research that examines students' problem representations. One potential avenue of research is exploring how students' problem representations differ when students are presented with pre-constructed problems (e.g., word problems) as opposed to situations where they pose their own problems (e.g., projects). As a first step in this endeavor, we examined high school students' representations of pre-constructed word problems dealing with hypothesis testing. We focused on these problems because (a) they are often used in statistics instruction, (b) the research on students' representations of such statistics problems is limited, and (c) we wanted to identify features students considered important in a controlled setting.

In statistics, students must be able to understand a problem so that they can apply statistical procedures appropriately and draw suitable conclusions. Knowing when to apply particular statistical procedures, such as hypothesis testing, is a difficult skill for students to acquire (Lovett, 2001; Quilici & Mayer, 1996). They must know the critical features that underlie statistical methods, recognize them in problems, and apply them appropriately, in order to be successful problem solvers (Hubbard, 1997). Recognizing structural features underlying methods such as a *t* test, chi-square test, and correlation is a necessary first step in deciding their appropriateness. This process is difficult for many students, who tend to rely on heuristics (e.g., a two-way table, hence the problem must require a chi-square test) rather than structural features (e.g., categorical data that examine a relationship between two variables, hence a chi-square test is required for this problem). Reliance on heuristics reflects a surface approach where key words or data structures are the focus for solving the problem. Emphasis on structural features reflects a principled approach where the purpose of analysis, its conditions of applicability, the type of data to be collected, the test algorithm, and the meaning of the conclusions are understood (Hubbard, 1997).

Statistics problems dealing with hypothesis tests can be represented in terms of various structural features. Higher level features for distinguishing between hypothesis tests involve making decisions about type of data (i.e., measurement vs. categorical), research purpose (i.e., examination of differences vs. relationships), and number of variables or groups (two vs. multiple; Howell, 1989). Finer grain features can include subcomponents of these higher level categories, such as measurement level (i.e., nominal vs. ordinal) and type of variable (i.e., related or dependent vs. independent). Research examining whether students represent statistics problems in terms of these features is limited. Some studies have investigated problem solving in the domain of statistics, focusing on problem-solving errors (e.g., Allwood, 1990; Allwood & Montgomery, 1982), problem-solving strategies and transfer (e.g., Paas, 1992), planning processes (e.g., Lovett, 2001), the relationship between competence and metacognition (Mariné & Enscribe, 1994), and expertise (Hauff & Fogarty, 1996; Hong & O'Neil, 1992). Only one study, by Quilici and Mayer (1996), has investigated students' problem representations.

Quilici and Mayer (1996) were interested in how college students with limited knowledge of statistics (i.e., only took an introductory statistics course) or no knowledge of statistics (i.e., never took a statistics course) represented inferential statistics problems. They examined students' problem representations after a session in which the students were exposed to examples that emphasized structure. This intervention was meant to help students recognize which tests were appropriate for which problems. The examples provided information about structure by identifying the inferential tests that were required to solve them, namely, *t* test, chi-square test, and correlation. Viewing these examples was expected to help students abstract the structural features underlying the three tests. The structural features emphasized in the problems consisted of type of variable (i.e., independent vs. dependent) and type of data (i.e., quantitative vs. categorical). After viewing the examples, students were required to group new problems that they thought belonged together. The effectiveness of the "structure-emphasizing" examples for fostering students' representations of inferential statistics problems was examined by comparing the problem groupings of students who received the intervention with the sorts of those who did not. Students who were shown sample statistics word problems for each test were expected to sort subsequent problems more on the basis of structure (i.e., group together all problems requiring the same test) than surface features (i.e., weather, politics, education) compared to students who were not exposed to examples. The results supported the hypothesis.

Quilici and Mayer's (1996) study was strong, but it was limited by its focus on a single measure, namely, scores indicating the extent to which problems were sorted based on structural and surface features. These data provided estimates of students' problem representations but did not reveal the thinking processes that formed the basis of the sorts. In other words, students were not required to explain *why* they sorted problems in the way that they did. Further research is needed to determine the validity of sorting scores as an indicator of problem representation. Moreover, additional sources of evidence such as verbal explanations are needed to triangulate the data and to provide insights into the reasons why particular problems were grouped.

The present study was exploratory and designed to extend Quilici and Mayer's (1996) work by collecting verbal data to examine the nature of student problem representations more closely. We were interested in why students sorted particular problems together and not just in how they were grouped. Our intent was to characterize students' problem representations immediately after receiving an instructional unit on hypothesis testing, rather than creating and examining the effectiveness of an intervention specifically designed to foster these skills. In this sense, the present study attempts to describe the problem representations that arise in typical statistics classrooms. Another way our study differed from Quilici and Mayer's was in the structural features that were emphasized in the sorting problems. The structure in our problems consisted of higher level features associated with making decisions about the appropriateness of inferential tests (i.e., purpose, data type, and number of variables or groups). Quilici and Mayer emphasized one of these higher level features (data type) and another general feature (variable type).

We replicated Quilici and Mayer's 1996 study only with respect to the sample, use of particular word problems (see methodology for details), the types of tests required to solve the problems (i.e., *t* test, chi-square test, and correlation), and the methodology employed (i.e., sorting task and scoring of problem groupings). Students participating in both studies were somewhat similar in that they were relatively comparable in terms of their statistical experience (however, Quilici and Mayer also included students who had no experience with statistics) and their ages (the difference between 12th grade and college is not substantial). Moreover, even though students had learned a range of tests, we chose to focus on *t* test, chi-square, and correlation problems for two reasons: (a) a desire to represent a range of tests frequently used in research and (b) time constraints. *Z*-test problems, for example, were not included because *Z* tests are rarely used in the analysis of most genuine research problems. Class instruction focuses on them because they provide the foundation for hypothesis testing and thus have conceptual value. Linear regression problems also were omitted because the distinction between regression and correlation was too subtle to be informative at this grade level. Finally, *F*-test problems would have been an interesting contrast to *t*-test problems. However, time constraints did not allow for their inclusion because students' participation would have been extended from one to two class periods, which was not possible due to students' schedules.

Three research questions were explored in this study:

1. Do the sorting scores indicate that problems are represented in a superficial or principled way?
2. Are the features identified in students' explanations based on statistical methods or irrelevant characteristics of problems?
3. Do students' explanations for sorts correspond with the scores representing the extent to which their sorts are superficial or principled?

Methods

Participants

Twenty-one students (11 female, 10 male) from a 12th-grade introductory statistics course participated in the study after completing instructional units on hypothesis testing in the spring. The sample was largely middle class and Caucasian (80%). Participants were among the strongest in their cohort in terms of mathematics ability. The content taught in the statistics course was equivalent to an introductory undergraduate-level course and was offered only to students who had successfully completed all required mathematics courses. The hypothesis testing units covered the following ordered topics: *Z* tests, *t* tests, chi-square tests, correlation, linear regression, and *F* tests. Students participating in the study thus had statistical knowledge and experience solving problems dealing with targeted inferential tests (i.e., *t* test, chi-square test, and correlation). Since each statistical method was learned in isolation from the others, there was little opportunity for students to participate in classroom activities that required they distinguish between the different tests.

Materials

Statistics word problems were typed and presented to students on separate 3-inch x 5-inch index cards. Problems represented three inferential tests students had learned about in class, that is, *t* test, chi-square test, and correlation. Three problems were developed for each test, resulting in a total of nine problems (adapted from Lovett, 2001; Quilici & Mayer, 1996). These problems can be represented in terms of two overarching features: structural and surface. Structural features can represent fundamental principles or statistical methods. In hypothesis testing, structural features can be inferential tests or characteristics associated with tests. Critical features underlying hypothesis tests include purpose of study (i.e., whether the research question involves examining differences or relationships) and data type (i.e., whether the data can be measured or counted). These two features form the basis of a principled problem representation in this context and provide the basis for deciding whether a chi-square test, a *t* test, or a correlation is appropriate for any given problem (Howell, 1989). Figure 1 illustrates features associated with each type of test.

An additional characteristic that could be salient for the students since they also learned about *F* tests is the number of groups or variables being compared. This feature enables students to distinguish a *t* test from an *F* test since they share the same purpose (i.e., comparison) and data type (i.e., measurement) features. A *t* test involves a comparison of two groups, whereas an *F* test is used to compare multiple groups. Consequently, a third feature, number of groups or variables, could be an

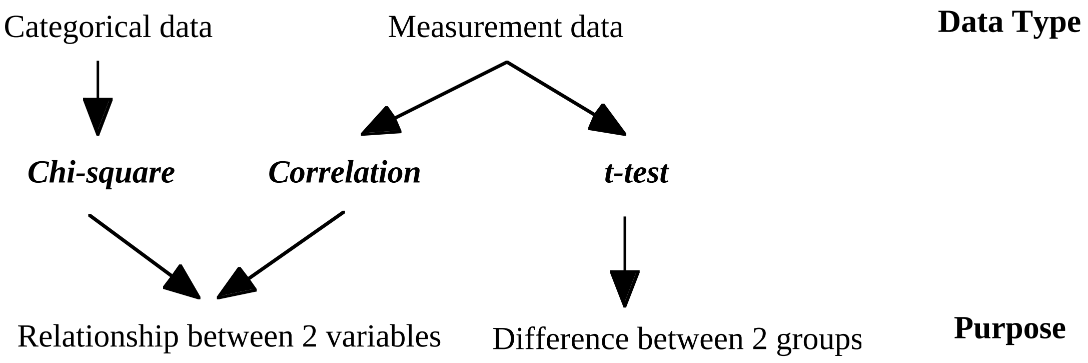

Figure 1. Critical features associated with *t* test, chi-square test, and correlation.

important part of the decision making for students on this task. It is therefore included as one of the salient categories even though the problems were not designed to vary in terms of this feature.

Surface features represent a superficial understanding of statistics and include a focus on semantic (i.e., topic or cover story, such as education, weather, and politics) or literal (i.e., data structure or organization) similarities. Table 1 illustrates how each problem varied by structure and surface features. The order in which problems were presented was counterbalanced to ensure that surface and structure problems were alternated. The resulting sequence was presented to each participant in the same order.

Procedure

Each student participating in the study was individually taken out of the statistics class to perform the sorting task. Before starting the task, participants were informed of what was required of them. They were then asked to explain what they had to do and how they would do it to ensure they understood the task. Students were given the deck of cards and instructed to sort problems based on how they "best went together" (Mariné & Enscribe, 1994; Quilici & Mayer, 1996). They were not required to solve the problems. Students also were required to think aloud and to explain the reasoning underlying each sort while they were performing the task. Several semi-structured interview questions were posed at the end of the task (e.g., How are problems in each pile similar? How is each pile different?). The interviewer recorded these sessions on audiotape and in a notebook. A blank worksheet was available for student use. The data were transcribed, segmented, and analyzed based on how problems were grouped together and why.

Measures

The data were examined in two ways. First, the manner in which problems were sorted was measured by computing scores that reflected the degree to which sorts were principled (i.e., structure score) or superficial (i.e., surface score). Second, the reasons (or explanations) underlying sorts were categorized according to structural (i.e., purpose, data type, number of variables or groups) or surface (e.g., cover story) features. The sorting scores and explanation categories are two indices of problem representation that vary in the amount of detail they provide. Each is described in more detail in Table 1.

Table 1

Statistics Problems Characterized by Structure and Surface

Structural features	Surface features (cover story)		
	Education	Weather	Politics
t test Difference Measurement 2 Groups	A professor is teaching two sections of the same class. One section meets on Mondays and Wednesdays, the other on Tuesdays and Thursdays. The professor gave the same test to both sections and wants to know whether students in the two sections performed differently. The test was worth a total of 100 points.	Weather reporters in the Pittsburgh area often give temperature readings that are based at two locations: the airport and the downtown core. A journalist wanted to find out whether the temperatures reported from the two locations varied. Temperature readings from both sites were recorded for one year.	A political candidate wants to know whether voters' party affiliation (Democrat vs. Republican) varies based on their income level. The candidate's aide conducts a survey asking people to report their party affiliation and their total annual income.
Chi-square test Relationship Categorical 2 Variables	A school superintendent suspects that high school students' intended college major varies by gender. To find out, a short questionnaire is distributed asking male and female senior students in the district whether they plan to major in the sciences or arts when they apply to college.	A weather analyst thought that there was a difference in the occurrence of tornadoes and hurricanes based on time of day. The scientist used data from the last 50 years that specified the type of wind phenomena and whether it occurred in the a.m. (i.e., midnight to noon) or p.m. (i.e., noon to midnight).	The governor's office wants to know if the prevalence of different kinds of crime varies across different regions of Pennsylvania. A state official collects crime reports from police stations across Pennsylvania. Each report is labeled with the name of the reporting police station and describes either a personal or a property crime.
Correlation Relationship Measurement 2 Variables	A professor teaching a class on creativity asks students to answer a questionnaire designed to measure creative thinking on a scale from 1-50. The professor believes that watching TV stifles creativity. Students' scores are recorded along with the reported number of hours of TV they watch per week.	After examining weather data for the last 50 years, a meteorologist claims that the annual precipitation varies with average temperature. For each of the 50 years, the meteorologist notes the annual rainfall and average temperature.	To receive additional federal funds to the health care budget, each state must obtain a government rating of the quality of its health care offerings (averaged across the state). A congressional aide wants to know whether the amount of federal funds allocated to each state depends on the state's health care ratings.

Sorts: Structure and surface scores. The degree of similarity between grouped problems was examined by computing a structure and surface score for each participant based on pairs of problems in each sort that represented structure and surface (Quilici & Mayer, 1996). Structure sorts were those in which *t*-test, chi-square, or correlation problems were correctly grouped together. Surface sorts were those in which education, politics, and weather problems were grouped together.

Four steps were involved in scoring the sorting data. First, structure and surface scores were calculated based on the way problems were paired. Three pairs can be made for each test type (e.g., *t*-test problem pairings = 1 & 2, 1 & 3, and 2 & 3) for a total of nine structure pairs (i.e., 3 *t*-test pairs, 3 chi-square pairs, and 3 correlation pairs). The same number of pairings applies to surface sorts where education, politics, and weather problems are grouped together. Optimal groupings thus consisted of three sorts with three problems each. Each pair was assigned a score from 0 to 3, for a maximum score of 9 for structure (i.e., 3 pairs for 3 types of tests) and a maximum score of 9 for surface (i.e., 3 pairs for 3 types of cover stories). Second, scores were converted into proportions by dividing each score by the maximum score. For example, a participant who produced four sorts consisting of two *t*-test problems, two correlation problems, two chi-square problems, and a mixed set of problems (i.e., one *t* test, one correlation, and one chi-square) would receive a structure score of 6. Dividing this number by 9 results in a proportion structure score of 0.67. Proportions ranged from 0 to 1. A proportion of 0 (minimum score) for structure or surface indicated that problems were not sorted based on principled or superficial features. A proportion of 1 (maximum score) indicated that the sorts were based on perfectly principled (i.e., for the structure score) or superficial (i.e., for the surface score) representations. The higher the proportion, the greater the degree to which the sort is based on structure or surface (i.e., maximum proportion score of 1 for both). The sort in the example would therefore be moderate in terms of structure and represent some principled understanding.

Third, participants were categorized as either structure- or surface-using, depending on the relative strength of each proportion. This measure emphasizes the predominantly stronger of the two problem representations and provides a general label. A participant was categorized as structure-using if the structure proportion was greater than or equal to the surface proportion. If the reverse was true, then the participant was characterized as surface-using. The fourth step in scoring the sorting data involved grouping structure proportions into categories to distinguish among three levels of problem representations: superficial, moderately principled, and principled. These categories provide a meaningful way of discussing the sophistication of the representations. Moreover, treating the data in this way allows for a comparison of scores with reasons underlying the sorts, which were categorized in this manner (as will be seen in the next section). We focused on structure proportions because they are the most interesting and account for the surface scores by virtue of the scoring method. High structure scores, for example, are generally associated with moderate to low surface scores and vice versa. Focusing on structure is also legitimate in cases where the structure and surface scores are equivalent. The structure proportions were grouped into representation levels in the following way: Proportions between 0 and 0.4 were categorized as superficial, proportions between 0.4 and 0.7 were coded as moderately principled, and proportions between 0.7 and 1 were identified as principled.

Reasons for sorts: Categories of features. Reasons for each sort were coded based on categories that emerged from the data. Many explanations reflected critical hypothesis testing features and included one or several of the following: (a) type of statistical test—*t* test, chi-square, correlation; (b) research purpose—difference versus relationship; (c) data type—measurement versus categorical; (d) number of variables or groups— two versus multiple; (e) superficial statistical considerations—data organization or whether the problem stated that data were collected; (f) nonstatistical considerations—topic or cover story; and (g) lack of problem similarity—problems not perceived as similar to others. The first three categories reflect structural features that underlie principled problem representations. The last three categories, on the other hand, are based on surface features that underlie superficial problem representations.

The categories were grouped in terms of sophistication levels to ascertain the degree to which explanations were principled. The data were grouped into principled, moderately principled, or superficial categories. A *principled* representation is reflected by explanations that are based on type of statistical test alone or in combination with underlying test features, such as purpose, data type, or number of variables or groups. Representing problems based on two or more of these critical features is also principled. Thinking about problems in these ways is sophisticated in that it requires that knowledge of multiple features be integrated. A *moderately principled* representation is based on explanations that consist of one critical feature. Considering a single feature is principled but incomplete because it is not connected to other critical features. A *superficial* representation is based on irrelevant considerations, such as data organization, whether the problem involved data that were already collected or whether data had to be collected to answer the question, cover story, or perceived lack of problem similarity. None of these features reflect statistical principles related to hypothesis testing. The percentage overlap between two raters independently coding the data was 87% before discussion for both the content and sophistication-level categories. Coding differences were resolved through discussion, resulting in 100% agreement.

Results

Students' problem representations were examined based on two types of data: sorting scores and explanation categories. These indices vary in the amount of detailed information that is provided about students' representations. First, we report on the nature and extent to which problem representations were principled based on students' explanations of their sorts. This index is more sensitive than sorting scores for identifying concepts students consider important and is therefore a more direct measure of problem representation. We focus on features that were identified by students and how sophisticated these are relative to key features involved in selecting appropriate hypothesis tests. These results are followed by an analysis of the average number of groupings that were produced and the extent to which problem representations were principled based on structure and surface scores. The score index is more sensitive to the degree of accuracy related to particular pairings of problems than to the knowledge associated with groupings. An accurate sort implies principled understanding. In this sense, a sorting score is an indirect measure of problem representation. We then compare the results obtained from the two measures to examine the relationship between sorting scores and explanation categories. We present cases in which the scores either underestimated or overestimated the degree to which problem representations were principled. Finally, we report on the types of problems that tended to be sorted together.

Sorting Categories

Explanations accompanying the sorts revealed that problem representations were highly variable. Twenty-six different reasons were generated. In some cases, one feature (e.g., purpose) provided the basis for each sort (58% of reasons) and in others, two or more features (e.g., purpose and data type) were focused on (42% of reasons). Features emphasized in students' explanations were grouped into 10 general categories. These categories and their relationship to each representation level are presented in Table 2.

Note that the data are based on the total number of groupings or piles that were produced (N = 72) rather than by student. Problems were generally not sorted according to an overarching criterion. Rather, students' reasons for sorting different groups of problems often were unrelated. For instance, problems were sorted into three piles because the student stated that one set of problems dealt with differences, another involved collecting data, and a third represented measurement data. In this case, the student was not using purpose as the overarching criterion for grouping problems. Otherwise two piles would have been made; one based on difference and the other on relationship. Instead, different kinds of reasons were emphasized for each of the three groupings, namely, purpose, superficial consideration, and data type. Variability of problem representation was therefore demonstrated in two ways: (a) the total number of different reasons provided and (b) the contrasting features that were emphasized for each sort that a student produced.

Table 2

Classification of Reasons for Sorts by Representation Levels

Representation level	% Reasons (N = 72)
Principled	23%
Purpose, data type, and number of variables / groups (n = 4)	6%
Purpose and data type (n = 6)	8%
Purpose and number of variables / groups (n = 1)	1%
Test and purpose (n = 5)	7%
Test (n = 1)	1%
Moderately principled	49%
Purpose (n = 24)	33%
Data type (n = 7)	10%
Number of variables / groups (n = 4)	6%
Superficial	28%
Superficial statistical (n = 13)	18%
Nonstatistical consideration (n = 4)	6%
No problem similarity (n = 3)	4%

Note. In situations where reasons emphasized both a critical feature (e.g., purpose) and an irrelevant feature (e.g., data organization), the strongest reason was prioritized and categorized accordingly (e.g., purpose).

The range of combined and single critical features in Table 2 illustrates that students' problem representations were based on key statistical ideas and varied in sophistication. Interestingly, purpose and data type, either in isolation or combination, formed the basis of most problem representations (59%), and were thus sufficient for sorting the problems in a principled way. Ability to think about statistics problems in terms of these critical features is important because it can enhance successful problem solving on word problems. Selecting an appropriate statistical test for solving statistical problems is based on one's ability to represent problems in terms of these features. While some explanations indicated that problems were represented in a principled way, half the representations were moderately principled, with purpose being the most salient single feature. Superficial representations also were evident but to a much lesser degree. Features such as data organization, which are based on statistics but are not particularly relevant to the task, formed the basis of most superficial groupings. Note that a focus on data organization reflects the often-used heuristic of data structure for determining a chi-square test. In short, students' explanations for each sort revealed that they had acquired sufficient statistical knowledge to enable them to represent the problems in terms of one or two critical test features.

Earlier we mentioned that students were not generally guided by an overarching criterion in grouping problems. As such, we examined the features that were emphasized in each sort, rather than by each student. One way to examine the extent to which each student's representation was principled is to average the levels that characterized each of the student's sorts. That is, sorts categorized as superficial, moderately principled, and principled were assigned scores of 0, 1, and 2, respectively. These scores were then averaged for each student (*N* = 21). For example, a student who made a superficial sort and two moderately principled sorts would be assigned a total score of 2, which divided by the number of sorts (i.e., 3), would be close enough to 1 to be categorized as moderately principled. Using this method, we found results similar to those reported above; that is, that most students provided reasons that reflected a moderately principled representation (57%). In addition, slightly more students provided reasons that reflected a principled (24%) rather than a superficial representation (19%).

Notice that we have not examined whether students were accurate in identifying the specific characteristics associated with each feature. For example, we did not determine whether a "relationship" explanation was correctly applied to a relationship problem. Rather, we focused on the general feature of purpose. The reason we did not code for accuracy of explanations was due to the ambiguity of students' responses. For example, a common explanation for sorts was that the problems dealt with "dependency." The problem with this response is it could mean any number of things. It could mean that the problem involves testing for a relationship or comparing paired groups. Students did not elaborate enough on their explanations to determine their accuracy with any degree of reliability. Hence, we decided to focus on whether students were able to think about statistics problems in terms of the more general features of purpose, data type, and number of variables or groups. Being able to do so is an important first step in developing the ability to select hypothesis tests appropriately.

Groupings and Sorting Scores

Students produced roughly the same number of groupings (*M* = 3.43, *SD* = 1.03) as would be expected from an optimal sort (i.e., 3). The actual problems that were grouped together, however, did not necessarily belong together. Two findings suggest that problems were sorted more in terms of statistical principles than superficial characteristics of the problems: (a) Students were characterized as being more structure-using (62%) than surface-using (14%)[2] and (b) the average structure proportion was higher (*M* = 0.454, *SD* = 0.268) than the average surface proportion (*M* = 0.243, *SD* = 0.160). Nonetheless, when we compared the average structure proportion to the optimal sort (i.e., proportion of 1), we found that problem representations were not substantially more principled than superficial. In fact, when the structure proportions were converted into representation levels, we found that most participants' sorts can be characterized as superficial (57%). Principled (24%) and moderately principled (19%) representations were evident but much less frequent. These findings contrast with those reported in the previous section and are discussed in more detail below.

Correspondence of Scores With Categories

The explanation categories suggested that performance was based on a variety of problem representations and that some of these were rather sophisticated. On the whole, this index indicated that students' representations tended to be moderately principled (57%). In this sense, students' explanations for their sorts were consistent

[2] Twenty-four percent of students were neither structure- nor surface-using since they received the same scores in both categories.

with the overall characterization of students' performance as structure-using. However, the degree of structure was more often low than high, with students' sorting scores being predominantly superficial (57%). Figure 2 illustrates that the two indices were nonetheless consistent in measuring principled problem representations, with 24% of the students demonstrating a strong understanding of hypothesis tests based on both their sorts and explanations. The inconsistent results at the moderate and superficial levels reflect the variability of explanations underlying the sorts and suggest that this measure is more sensitive than scoring based on actual performance (i.e., sorts produced).

Table 3 illustrates the limited relationship between sort scores and explanation categories in identifying the degree to which students' ($N = 21$) problem representations were principled. Only 5% of the overall principled representations were actually principled in terms of both structure scores and explanations. The only real consistency occurred for superficial representations where both sorts and explanations were characterized as superficial. Surprisingly, there were quite a few superficial-scoring students who were principled in their explanation (14%). The

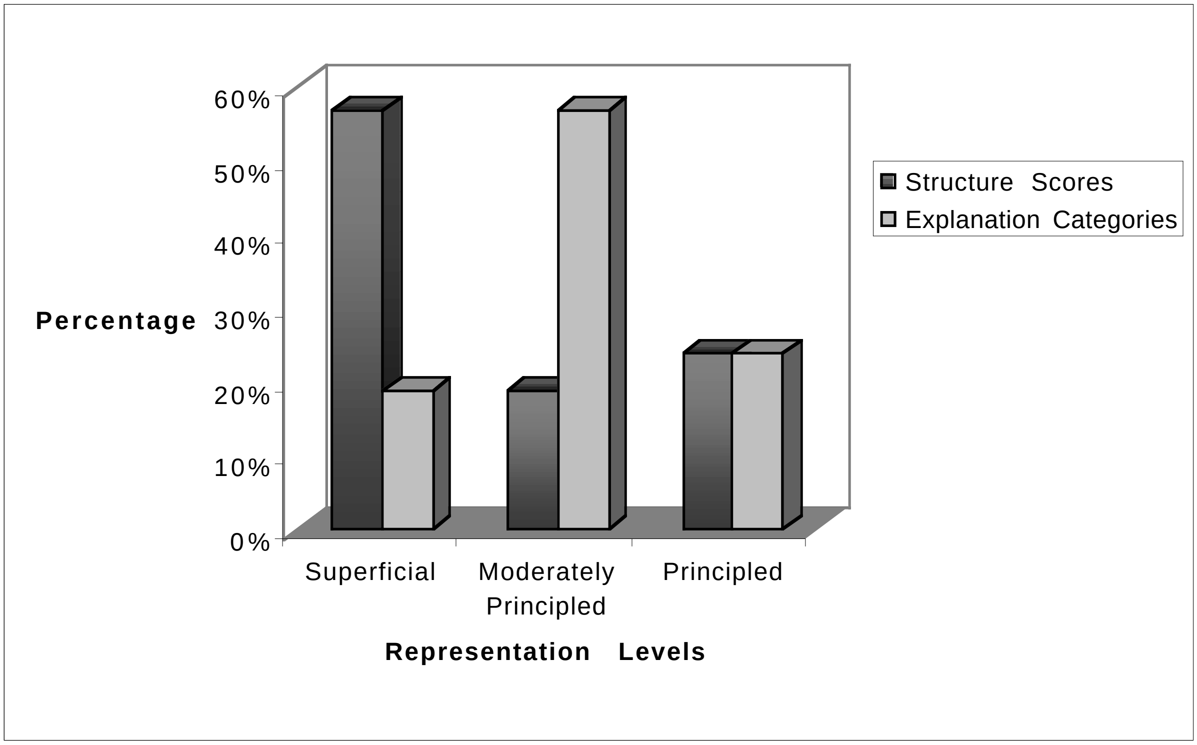

Figure 2. Representation level by index.

Table 3

Relationship Between Two Indices of Problem Representation

	Explanations			
Structure score	Superficial	Moderately principled	Principled	Total
Superficial	14% (*n* = 3)	29% (*n* = 6)	14% (*n* = 3)	57% (*n* = 12)
Moderately principled	5% (*n* = 1)	9% (*n* = 2)	5% (*n* = 1)	19% (*n* = 4)
Principled	0% (*n* = 0)	19% (*n* = 4)	5% (*n* = 1)	24% (*n* = 5)
Total	19% (*n* = 4)	57% (*n* = 12)	24% (*n* = 5)	100% (*N* = 21)

reverse did not occur. Superficial reasons did not underlie principled sorts. These data highlight the variability of students' representations and suggest that students understand more about statistics than might be implied by performance alone. Weak performance, as indicated by sorts, may, in fact, underestimate students' level of understanding, which is revealed more explicitly in students' explanations of their sorts. Students may be able to represent problems in principled ways by emphasizing critical features that underlie inferential tests. However, they may not be able to attribute features to the correct problem. It is this connection or integration problem that must be addressed in instruction and assessments.

Two additional examples illustrate the variability of representations and the potential of either underestimating or overestimating students' knowledge. The first example involves a case where the same level of performance was exhibited for entirely different reasons. The clearest example from the data involves four students whose sorts were scored as principled (i.e., same structure proportion score of 0.8), but whose reasons varied in sophistication among sorted problems and between students. Table 4 displays the reasons underlying these students' sorts. Three points can be made regarding these data. First, some criteria are guiding the sorts, but these are not consistently applied to all problems (e.g., student C distinguishes two sets of problems based on whether or not data were collected but sorts the remaining problems based on the assumption that differences are being compared). Second, one would expect the explanations to reflect primarily principled representations or at least a mix of principled and moderately principled representations given the high structure score. Yet, there were more moderately principled and superficial reasons underlying these principled sort scores than there were principled explanations. Third, reasons provided for the sorts varied greatly by student. The

Table 4

An Example of Different Reasons Underlying a Principled Sort Score

Student A	Student B	Student C	Student D
Principled: Difference + measurement	*Moderate*: No. of groups	*Moderate*: Difference	*Moderate*: Difference
Moderate: Categorical	*Moderate*: No. of groups	*Superficial*: Data were collected	*Moderate*: Relationship
Superficial: Data organization	*Moderate*: No. of groups	*Superficial*: Data were not collected	

only feature that seemed salient to most of these students is associated with purpose, that is, difference. In this example, students' understanding of inferential statistics based on their sort scores was overestimated. Even though students performed well on the task (i.e., tended to sort the appropriate problems together), their reasons for doing so varied and were not always principled.

A second example illustrating variability of representations involves a case where the same problems were grouped together for different reasons. Four students correctly sorted correlation problems together, receiving a perfect structure score for this problem set. Two students were principled in their reasoning, one indicating that the problems involved correlation and the other that they involved making a comparison of averages. A third student demonstrated a moderately principled representation by focusing on number of groups (i.e., multiple options for both variables). The fourth student, also moderately principled in representing these correlation problems, focused on purpose (i.e., one independent variable having an effect on another). Again, reasons varying in sophistication and content can underlie strong performance. These examples illustrate how seemingly similar performance can mask variability in thinking and how perfect scores (or seemingly correct answers) do not necessarily reflect principled understanding

Types of Problems Sorted Together

The previous sections presented results pertaining to students' problem representations based on two kinds of measures: scores derived from students' sorts and categories derived from explanations. We presented data suggesting that these two measures did not necessarily provide consistent information about the degree to which problem representations were principled. In this section, we focus on the types of problems that tended to be sorted together to gain insight into the confusions that may affect students' performance on the sorting task. We examined the frequency with which certain types of problems were grouped together and found that confusion was strongest for chi-square problems. Chi-square problems tended to be paired with either *t*-test (42%) or correlation (39%) problems. According to students' explanations, these problems were grouped together mainly because they involved an examination of differences and dependency (closely related to relationship), respectively. *T*-test and correlation problems occasionally were grouped together mostly because of their common data type (19%). These results suggest that students focused on the correct features, but had difficulty distinguishing among problems because their representations were not sufficiently integrated. For example, sorting *t*-test and correlation problems together because they share the same data type is accurate. However, the only way to distinguish between these two tests is to go one step further and decide whether the problems examined differences or relationships. Considering both purpose and data type is required for distinguishing between problems requiring these tests (or for any of the problems presented in this study).

Summary

Two main findings emerged from the data in this study. The first was that students' problem representations were based on important statistical features, and the extent to which they were emphasized in students' explanations was variable. Most students were able to think of the statistical problems in terms of purpose, that is, whether they involved an examination of differences or relationships. Other students were able to think in terms of data type or number of variables or groups. A smaller number of students thought of the problems in terms of both purpose and data type. Although not many students focused on superficial characteristics of the problems, some did think about problems in terms of both superficial and principled features (or moderately principled). These mixed representations reflect students' incomplete understanding of inferential tests. The lack of knowledge integration also was seen in the types of problems that were grouped together. By focusing on a single correct feature, students grouped two sets of problems that did not belong together. Students would have been able to tease apart these problem sets had they considered the critical second feature.

The second main finding of the study was that students' explanations were much more sensitive to the nature of their problem representations than was their sorting performance. The specific features guiding the performance were revealed through explanations, and in many cases, the performances were either underestimated or overestimated by the sorting scores. The sorts suggested that students' problem representations were predominantly superficial, whereas the explanations revealed that the representations were in fact more moderately principled than superficial. In this case, students were able to think about the important features but were not successful in applying them to the appropriate problems. The sorting scores alone would have underestimated the extent to which students' representations were principled. However, there was also the case in which performance overestimated students' problem representations. The scores suggested a principled representation, whereas explanations revealed that a mix of features formed the basis of the sorts. These features, however, were not always principled and tended to be moderately principled or a mix of all three types of representations. These findings suggest that assessments that provide students with an opportunity to explain their thinking are most sensitive to capturing the strengths and weaknesses in their understanding of statistical methods.

Discussion

Problem representation is a fundamental aspect of problem solving and has been investigated extensively in the domains of physics and mathematics. The literature reveals that successful problem solvers represent problems they are attempting to solve in terms of structural features that are critical to solving problems within that domain. In contrast, unsuccessful problem solvers represent problems in terms of superficial features, which are irrelevant to solving the problem. Problem representation becomes more important with increased use of problem- and project-based activities in K–12 classrooms. In these activities, students can pose their own questions or problems to investigate, which increases the ill-defined nature of such tasks. Moreover, the inclusion of statistics instruction at this level and the limited problem-solving research in this area enhance the value of studies examining students' representations of problems constructed by themselves or others. This exploratory study was a first step in investigating students' problem representations in statistics by focusing on word problems constructed by researchers.

The 12th-grade students who participated in the study had a strong background in mathematics and represent young adults who take or prepare for the Advanced Placement (AP) statistics courses. Although these students are members of a select group, the challenges they encountered can help statistics instructors focus on issues that are likely to arise for learners of other ability levels. In other words, the difficulties encountered by the higher ability students in our sample are likely to occur for all students. Moreover, although hypothesis testing currently is not a standard statistical unit in the high school mathematics curriculum, it is likely to become more mainstream as students' sophistication with statistics increases due to their early experiences with this content. Thus, this exploratory study can provide us with a glimpse of what needs to be done for high school students who are currently learning inferential statistics and for building this understanding in the earlier grades. In this sense, the results of the study raise important issues that can then be tested in the general population, as hypothesis testing becomes a more common unit in mathematics or even in science.

The findings from our exploratory study reveal that the 12th-grade students were able to develop somewhat deep structural representations even without direct instruction on underlying test features or on how to select an appropriate test. This finding counters Quilici and Mayer's (1996) contention that students are predisposed to represent problems superficially until they receive an instructional intervention, such as structure-emphasizing examples, that reduces this tendency. The contrasting results could be due to one reason or a combination of reasons: (a) the collection of verbal data in our study; (b) the instruction; and/or (c) the limited representativeness of our sample. A likely explanation for differences in the two studies is the use of verbal data in our study, which was missing in Quilici and Mayer's study. If we had focused solely on sorting performance, the results of our study would have concurred with Quilici and Mayer's contention that students tend to sort problems superficially. However, the verbal data in our study revealed that despite sorting problems in a superficial way, students' explanations for these sorts were moderately principled; they focused on at least one of three critical features. On the basis of these verbal data, problem representations were moderately principled rather than superficial.

A second explanation for the contrasting result is the manner in which students learned statistics initially. Little information is provided about the initial classroom experience of Quilici and Mayer's (1996) sample except that some participants had experience with statistics through an introductory course. The instruction in these college courses still tends to be somewhat traditional. Students participating in our study, however, learned statistical content by building models, and the curriculum was developed based on a variety of sources, including AP statistics. The initial foundation for our students' statistical knowledge was therefore based on instruction that was likely more innovative than traditional. Note that although the teacher's approach was innovative, little time was available for emphasizing the relationship among tests and for explicitly addressing how they contrast in terms of the underlying features. In this sense, direct instruction on underlying test features or on how to select an appropriate test was not provided. Only in this way was the initial instruction in both studies somewhat similar. However, Quilici and Mayer's sample did subsequently receive instruction for emphasizing structure, albeit indirect. Students in our study did not receive such instruction.

A third source of differences between the two studies that might account for the results is the composition of our sample in terms of ability level. Quilici and Mayer (1996) found that their high-ability college students always sorted statistics problems based on structure, regardless of whether they were shown an example word problem. Our sample can be regarded as being predominantly high ability since students were among the most successful in mathematics in their cohort. As such, our sample may have been predisposed to seeing more structure than most of the students participating in Quilici and Mayer's study. Moreover, some participants in their study did not have any previous statistics experience and were thus less knowledgeable initially than other participants in both their study and ours who had taken an introductory statistics course. It is also possible that we created a demand characteristic by asking students to perform a task during their statistics class that predisposed them to think about the problems in a statistical manner.

Nonetheless, in our study, there was clearly room for improvement in students' sorting performance, which was not optimal. Quilici and Mayer (1996) also found that sorting performance was not at a high level despite improvement after exposure to structure-emphasizing examples. Variability in problem representation is a possible source of the less-than-optimal performance. In our study, problem representations were highly variable. This finding is not surprising since students had acquired only a moderate level of experience. As apprentices (Collins, Brown, & Newman, 1989), students' initial representations were based on different forms of knowing that co-exist and may even be contradictory (Siegler, 1996). The statistics course was their first, and although they acquired some statistical knowledge, they had not yet mastered the conditions under which this knowledge applied. Students needed more experience applying their knowledge in a variety of contexts (Silver & Marshall, 1990) and on different kinds of questions (Hubbard, 1997). Performance in this moderate learning phase is usually diverse and fragmented.

Part of the variability in students' problem representations is due to a lack of knowledge integration. The expertise literature demonstrates that knowledge of concepts, methods, and principles in a domain becomes increasingly interconnected as competence is achieved (Glaser, 1989). These connections reflect an organized knowledge base that forms the basis of principled representations. In contrast, novices' knowledge is fragmented, and their understanding superficial. In this study, students tended to represent problems in terms of one critical test feature, that is, purpose of study. Problems requiring different types of tests were confused because they had a single characteristic in common (e.g., purpose). The second feature that distinguished problem sets (e.g., data type) was often ignored. Students must be able to make connections among concepts and between ideas and skills in order to succeed in solving statistics problems (Hauff & Fogarty, 1996; Huberty, Dresden, & Bak, 1993; Schau & Mattern, 1997). The finding that our sample's knowledge was not sufficiently integrated reflects the fact that they were early learners of inferential statistics.

Educational Implications

The instructional implications that can be drawn from this study involve a focus on experimental design, a change in the sequencing of content, and the use of examples. First, statistics courses dealing with inferential statistics could benefit from an emphasis on experimental design rather than on the mastery of computation for using test formulas (Quilici & Mayer, 1996). Students participating in our study were able to detect at least one critical test feature from a single course on hypothesis testing. This is a positive finding. Unfortunately, the connection among structural features and how they distinguish between hypothesis tests was not made. This difficulty can be attributed to the teaching of each hypothesis test separately (Lovett, 2001; Lovett & Greenhouse, 2000). Opportunities for comparing and contrasting hypothesis tests in statistics classrooms are rarely provided. Moreover, structural features that underlie statistical methods are not explicitly addressed. Many courses focus on the computation of statistics for each test; few emphasize the critical features underlying each, and when they do, only one lesson is provided on the topic. The sorts of students in Quilici and Mayer's study were not optimal partly because test features were emphasized indirectly. Students had to abstract the features from examples that identified the test, rather than explicitly highlight the critical features. Lovett found that students without prior statistics experience improved significantly in their problem sorts (i.e., from superficial to structural) when the instruction highlighted the structural features and their meaning. Teachers therefore need to provide students with opportunities to examine word problems in terms of experimental design features. They should be asked to specify the variables and their characteristics in each problem (Quilici & Mayer). A focus on differences in data types for dependent variables is especially needed given confusions between the *t* test and chi-square test on the one hand (Quilici & Mayer), and the chi-square test and correlation on the other. The connection between purpose and data type can be emphasized each time a hypothesis test is taught, thereby providing contrasts between tests and facilitating the development of an organized structure of knowledge.

Increased focus on features associated with experimental design can be accomplished by changing the sequence in which the content associated with inferential statistics is taught. In mathematics and statistics classrooms, students spend most of their time learning equations and formulas, which they are exposed to first, and then are given opportunities to solve word problems at the end of each unit. Some researchers recently have suggested that the sequence should be reversed, that is, students should be given experience with word problems first and then be required to solve equation or symbol manipulation problems (Brenner et al., 1997; Koedinger & Nathan, 1999; Nathan & Koedinger, 2000). The key is for students to understand when it is appropriate to apply a particular method before trying to solve a problem using formulas. It is the decision-making aspect of this process that is difficult for students to grasp. Koedinger and his colleague (Koedinger & Nathan; Nathan & Koedinger) have suggested that instruction should build on students' representations of simple word problems, which are based in natural language. Verbal representations are more familiar to learners than formal representational systems, which can be more difficult to understand in some cases.

In statistics, beginning instruction with word problems can provide teachers with an opportunity to emphasize the structural features associated with each test at the start of each section covering a different test. These features could be taught cumulatively so that similarities and differences in features across inferential tests are emphasized throughout the curriculum. One instructional strategy that seems to help individuals to perceive the structure across problems is the use of examples. Examples appear to be useful under the following conditions: (a) when differences between pairs of problems are highlighted (Catrambone & Holyoak, 1989; Gick & Holyoak, 1983); (b) when multiple examples are used with high-ability students (Sweller & Cooper, 1985); and (c) when example solutions are modified to highlight the use of principles (Catrambone, 1994). Presenting students with examples that emphasize structure before sorting or solving word problems also seems to improve students' ability to represent problems based on structure (Paas, 1992; Quilici & Mayer, 1996; Silver & Marshall, 1990). To ensure that problems are represented in a principled way, the features that underlie specific tests should be addressed explicitly.

Variability in students' problem representations and the different level of detail provided by the two indices used in this study have a clear implication for assessment, namely, that multiple forms of assessment are needed to validly assess aspects of problem solving. A full portrayal of students' problem representation would not have been achieved without the use of multiple forms of evidence. Jacobs (1993) found that paper-and-pencil assessments underestimated students' understanding compared to assessments that provided students with an opportunity to explain their thinking. In a similar vein, requiring that students explain their rationale for sorting certain problems together enabled us to notice some inconsistencies between problem groupings and the explanations. Students' performance on the sorting task both underestimated and overestimated their understanding. The consequence of this variability for assessment is that some labels for describing the nature of representations indicate that students could solve the problem (e.g., scores), while others suggest they could encounter difficulties (e.g., explanations), or vice versa. Using an index that illustrates the degree to which representations are principled by way of problem similarity is sensible. This index provides a quick and efficient way to examine students' representations and to get a sense of their underlying knowledge structures. However, this measure is not completely precise, as demonstrated in this paper. This finding raises the question of how much precision is required for a measure to be useful. In this case, precision was enhanced with explanations, thereby increasing the utility of representations for judging poor and strong problem solving.

Multiple forms of assessment can provide a more complete and detailed profile of student learning, as well as increase validity (Collins, Hawkins, & Frederiksen, 1993-1994; Costa, 1989; Frederiksen & Collins, 1989; Huberty et al., 1993; Lajoie, 1995; Linn, Baker, & Dunbar, 1991; NCTM, 1995; Shepard, 1989, 1991; Wiggins, 1990, 1992). The potential for misrepresenting what students actually know and do not know on a single assessment is lessened when multiple forms of assessments are used. Hubbard (1997) suggested that assessments include a variety of problems and that at least one of these be nonstandard. Students should also be encouraged to apply their knowledge in new ways. An alternative assessment could require that students pose their own problems rather than solve pre-constructed ones (Hubbard, 1997). This suggestion is consistent with the notion of having students investigate projects or problems of their own choosing. It would be interesting to examine students' problem representations within this context. Here, too, students eventually would have to consider the essential features of the problems for selection of an appropriate statistical analysis for the project data.

The sorting task has typically been used as a methodology within psychology, rather than as a form of assessment within education. Nonetheless, this task has value in statistics classrooms as a formative assessment for improving ongoing learning and instruction. It is a short task (students took, on average, 13 minutes to complete it) that can be accomplished individually or within small groups. The key is to have students *explain* their problem groupings on paper and/or in group or class discussions. Using this task as an assessment is a quick, efficient, and valuable way to receive feedback on statistical content students perceive as important, as well as about what they ignore or do not fully understand. Students can focus on a variety of features and at different grain sizes. The sorting task can highlight the extent of students' knowledge as long as it is accompanied by explanations. The teacher could use this information to revise the instruction to focus on relevant concepts and to foster knowledge integration. This form of assessment is dynamic and can be embedded within the instruction.

In conclusion, multiple and formative forms of assessments are needed to obtain valid information about various aspects of students' thinking and performance during the problem-solving process. Problem representation continues to be a critical component of problem solving. We have presented data to suggest that although students can perform at a particular level, the thinking that underlies the performance can be superior. Instructional activities and assessment tasks must provide learners with opportunities to articulate what they consider important and to identify similarities and differences among solution methods. Integrating assessment with instruction so that one builds on the other in this manner is the hallmark of formative assessment.

References

Allwood, C. M. (1990). On the relation between justification of solution method and correctness of solution in statistical problem solving. *Scandinavian Journal of Psychology, 31,* 181-190.

Allwood, C. M., & Montgomery, H. (1982). Detection of errors in statistical problem solving. *Scandinavian Journal of Psychology, 23,* 131-139.

Becker, B. J. (1996). A look at the literature (and other resources) on teaching statistics. *Journal of Educational and Behavioral Statistics, 21* (1), 71-90.

Brenner, M. E., Mayer, R. E., Moseley, B., Brar, T., Durán, R., Reed, B. S., et al. (1997). Learning by understanding: The role of multiple representations in learning algebra. *American Educational Research Journal, 34,* 663-689.

Catrambone, R. (1994). Improving examples to improve transfer to novel problems. *Memory and Cognition, 22,* 606-615.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 1147-1156.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55-81.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, A., Hawkins, J., & Frederiksen, J. R. (1993-1994). Three different views of students: The role of technology in assessing student performance. *Journal of the Learning Sciences, 3,* 205-217.

Costa, A. L. (1989). Re-assessing assessment. *Educational Leadership, 46*(7), 1.

Cumming, G., Thomason, N., & Zangari, M. (1995). Designing software for cognitive change: StatPlay and understanding statistics. In J. D. Tinsley & T. J. van Weert (Eds.), *World Conference on Computers in Education (WCCE) IV, Liberating the learner* (pp. 753-765). London: Chapman & Hall.

Derry, S. J., Levin, J. R., Osana, H. P., & Jones, M. S. (1998). Developing middle-school students' statistical reasoning abilities through simulation gaming. In S. P. Lajoie (Ed.), *Reflections on statistics: Agendas for learning, teaching, and assessment in K-12* (pp. 175-195). Mahwah, NJ: Lawrence Erlbaum Associates.

Fillebrown, S. (1994). Using projects in an elementary statistics course for non-science majors. *Journal of Statistics Education* [Online], *2*(2). http://www.stat.ncsu.edu/info/jse/v2n2/fillebrown.html

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*, 44-63.

Gick, M., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 12*, 306-355.

Glaser, R. (1989). Expertise and learning: How do we think about instructional processes now that we have discovered knowledge structures? In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 269-281). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hauff, H. M., & Fogarty, G. J. (1996). Analysing problem solving behaviour of successful and unsuccessful statistics students. *Instructional Science, 24*, 397-409.

Hong, E., & O'Neil, H. F., Jr. (1992). Instructional strategies to help learners build relevant mental models in inferential statistics. *Journal of Educational Psychology, 84*, 150-159.

Howell, D. C. (1989). *Fundamental statistics for the behavioral sciences* (2nd ed.). Boston: PWS-KENT Publishing Co.

Hubbard, R. (1997). Assessment and the process of learning statistics. *Journal of Statistics Education* [Online], *5*(1). http://www.stat.ncsu.edu/info/jse/v5n1/hubbard.html

Huberty, C. J., Dresden, J., & Bak, B-G. (1993). Relations among dimensions of statistical knowledge. *Educational and Psychological Measurement, 53*, 523-532.

Jacobs, V. R. (1993). *Stochastics in middle school: An exploration of students' informal knowledge*. Unpublished master's thesis, University of Wisconsin, Madison.

Koedinger, K. R., & Nathan, M. J. (1999). *The real story behind story problems: Effects of representations on quantitative reasoning.* Manuscript submitted for publication, Carnegie Mellon University/University of Colorado.

Lajoie, S. P. (1995). A framework for authentic assessment in mathematics. In T. A. Romberg (Ed.), *Reform in school mathematics and authentic assessment* (pp. 19-37). Albany: The State University of New York (SUNY) Press.

Lajoie, S. P., Lavigne, N. C., Munsie, S. D., & Wilkie, T. V. (1998). Monitoring student progress in statistics. In S. P. Lajoie (Ed.), *Reflections on statistics: Agendas for learning, teaching, and assessment in K-12* (pp. 199-231). Mahwah, NJ: Lawrence Erlbaum Associates.

Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Models of competence in solving physics problems. *Cognitive Science, 4*, 317-345.

Lavigne, N. C., & Lajoie, S. P. (2000). *Statistical reasoning in investigations: The case of problem posing.* Manuscript submitted for publication, University of Pittsburgh / McGill University.

Lester, F. K. (1994). Musings about mathematical problem-solving research: 1970-1994. *Journal for Research in Mathematics Education, 25*, 660-675.

Linn, R. L., Baker, E L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Lovett, M. C. (2001). Collaborative convergence on studying reasoning processes: A case study of statistics. In S. M. Carver & D. Klahr (Eds.), *Cognition and Instruction: 25 Years of Progress* (pp. 347-384). Mahwah, NJ: Lawrence Erlbaum Associates.

Lovett, M. C., & Greenhouse, J. B. (2000). Applying cognitive theory to statistics instruction. *The American Statistician, 54*(3), 1-11.

Mariné, C., & Enscribe, C. (1994). Metacognition and competence on statistical problems. *Psychological Reports, 75*, 1403-1408.

Marshall, S. P. (1995). *Schemas in problem solving*. Cambridge, MA: Cambridge University Press.

Mayer, R. E. (1985). Implications of cognitive psychology for instruction in mathematical problem solving. In E. A. Silver (Ed.), *Teaching and learning mathematical problem solving: Multiple research perspectives* (pp. 123-138). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mayer, R. E., Lewis, A. B., & Hegarty, M. (1992). Mathematical misunderstandings: Qualitative reasoning about quantitative problems. In. J. I. D. Campbell (Ed.), *The nature and origins of mathematical skills* (pp. 137-153). North Holland, Amsterdam: Elsevier Science Publishers B. V.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65*, 123-165.

Nathan, M. J., & Koedinger, K. R. (2000). Teachers' and researchers' beliefs about the development of algebraic reasoning. *Journal for Research in Mathematics Education, 31*, 168-190.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Paas, F. G. W. C. (1992). Training strategies for attaining transfer for problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84,* 429-434.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88,* 144-161.

Schau, C., & Mattern, N. (1997). Assessing students' connected understanding of statistical relationships. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 91-104). Amsterdam: IOS Press.

Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 484-494.

Shaughnessy, M. J. (1996). Emerging research issues in the teaching and learning of probability and statistics. In B. Phillips (Ed.), *Statistical education* (pp. 39-48). Swinburne, Australia: International Association for Statistical Education.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership, 46*(7), 4-9.

Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20*(7), 2-16.

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking.* New York: Oxford University Press.

Silver, E. A. (1981). Recall of mathematical problem information: Solving related problems. *Journal for Research in Mathematics Education, 12,* 54-64.

Silver, E. A., & Marshall, S. P. (1990). Mathematical and scientific problem solving: Findings, issues, and instructional implications. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive thinking* (pp. 265-290). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sweller, J., & Cooper, G. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2,* 59-89.

Wiggins, G. (1990). *The case for authentic assessment*. (Contract No. R-88-062003). Washington, DC: Office of Educational Research and Improvement. (ERIC Document Reproductive Service No. ED 328 611).

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership, 49*(8), 26-33.