Assessing Expert Knowledge Representations of Introductory Statistics

CSE Technical Report 600

Robert Glaser
CRESST, LRDC / University of Pittsburgh

June 2003

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 2.4 Classroom and Teacher's Assessment
Robert Glaser, Project Director, CRESST, LRDC / University of Pittsburgh

The work reported herein was supported under the Educational Research and Development Centers Program, PR / Award Number R305B960002-01, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

ASSESSING EXPERT KNOWLEDGE REPRESENTATIONS OF INTRODUCTORY STATISTICS

Nancy C. Lavigne and Robert Glaser
LRDC/University of Pittsburgh

Abstract

The assessment challenge today is to build assessments based on empirical and theoretical knowledge of learning and cognition. A cognitive proficiency that has received relatively little attention in statistics classrooms is how students represent their knowledge of statistical content. In this report, we focus on two assessments that are specifically designed to assess different but related aspects of individuals' knowledge representations: (a) problem sorts which measure how individuals represent a specific aspect of their knowledge in the context of word problems (i.e., problem representation) and (b) concept maps which measure how individuals represent their knowledge of the discipline as a whole (i.e., domain representation). This case study is meant to assess a statistics expert's representation of statistics as explicitly as possible. The participant was required to (a) group problems based on how he felt they best belonged together and explain his sorts after he was finished, (b) identify and explain various levels of problem groupings he expected introductory statistics students to generate, and (c) construct two concept maps, one representing statistics as a whole (includes descriptive and inferential statistics) and a second representing a specific aspect of statistics (inferential statistics), which was content emphasized in the problem sorting task. Explanations were audiotaped. The data suggest that the problem sorting task can be useful as a measure of representation when supplemented with an individual's explanations for his or her sorts. Moreover, concept maps can assist in the interpretation of performance on the problem sorting task. The concept map and problem sorting measures can be mutually informative, with the concept map providing a broader picture and the problem sorts illustrating how certain concepts become salient when applied to different contexts.

Researchers have made significant strides in developing an understanding of the learning and cognition demonstrated by students at various levels of education. While it is true that many questions remain to be investigated, a substantial body of knowledge has been acquired about how individuals learn academic content, such as mathematics and science. This body of work, in turn, has led to a progression in and expansion of views, at least within the research community, about learning and performance. Unfortunately, the growing knowledge base and perspectives regarding learning and cognition are not always reflected in assessment practices.

Consequently, what is assessed is often unrepresentative of the range of cognitive abilities that individuals can demonstrate (Glaser, 1990; Wolf, Bixby, Glenn, & Gardner, 1991). According to the National Research Council (NRC, 2001) and some researchers interested in assessment (e.g., Shepard, 1991; Wiggins, 1993), a fundamental problem in education today is that typical assessments of achievement are based on traditional beliefs (e.g., thought is composed of independent pieces of knowledge and skills that are acquired by passively receiving information) that are inconsistent with current perspectives on cognition and learning (e.g., thought is composed of connected chunks of knowledge that are acquired by actively trying to make sense of information). Current views consider the cognitive processes involved in learning to be an important factor in explaining how learners become proficient. The extent to which certain processes are emphasized, however, differs across perspectives. Some approaches focus more on external factors, such as social interaction and culture (e.g., social perspectives); some emphasize internal factors (e.g., schema theory); and others attempt to integrate these views by considering both internal and external factors (e.g., situated cognition) (Reynolds, Sinatra, & Jetton, 1996).

In this paper, we are concerned with a critical aspect of problem solving, namely, problem representation. We are particularly interested in examining how individuals who are experienced in statistics represent their knowledge of statistics in the context of word problems. This work is based on a previous study that suggested that measures more direct than problem sorts were needed to validly assess the nature of problem representations (Lavigne & Glaser, 2001). We therefore examine representations in a more direct way by examining an individual's representation on a concept mapping task in addition to the problem sorting task.

The view that is clearly associated with individuals' knowledge representation is schema theory. We refer to this theory for the following reasons: (a) the tasks described in this paper are meant to be performed individually where the social context has a limited role, and (b) the concern is with individuals' knowledge at a specific stage of understanding rather than with the development of that knowledge and the factors that influenced its growth. From the perspective of schema theory, individuals learn by trying to make sense of information they encounter. Knowledge is constructed by building on what is already known and by transforming information in a meaningful way (McKendree, Small, Stenning, & Conlon, 2002). Knowledge is organized into structures referred to as schemata. In problem solving

situations, schemata result from experience in solving problems that share common underlying features that are meaningful in that domain (Marshall, 1995). The most relevant features are abstracted and incorporated into existing structures or form the basis of new schemata. In this sense, a schema is a representation of the knowledge that learners acquire, construct, and organize. As learners gain experience, their knowledge becomes integrated and organized such that the concepts in a domain and their interrelationships are structured around key principles or ideas (Chi, Feltovich, & Glaser, 1981; Larkin, McDermott, Simon, & Simon, 1980). Representing knowledge in this way (i.e., in terms of principles) facilitates recall and results in competent performance. How learners represent their knowledge of a domain and how changes in their representations lead to competence is a hallmark of expertise. However, typical assessments ignore this fundamental aspect of competence. Multiple-choice tests are not generally designed to assess students' understanding of the relationships between concepts in a content area (Schau & Mattern, 1997).

According to NRC (2001), assessments should be designed based on a model of cognition and learning that relies on the most current understanding of how learners represent their knowledge and develop competence in particular content areas. The research described in this paper is intended to address this omission by examining how experts in the domain of statistics represent their knowledge. Statistics education is a domain worthy of investigation because it is a young discipline (Garfield, 2002) and has recently become an area of interest both pedagogically and empirically. The visibility of statistics has increased with its addition to the K-12 mathematics curriculum and with use of inquiry-based activities in science classrooms that require drawing meaningful conclusions from data. Similarly, statistics educators at the university level are becoming more focused on investing in research that informs their practice (Jolliffe, 2002). The research conducted in this area thus far suggests that undergraduate students perform worse on test items that are designed to measure conceptual understanding (i.e., require knowledge of more than two concepts and their relationship) than on items evaluating computational skills (Huberty, Dresden, & Bak, 1993; Kottke, 2000). Schau and Mattern (1997) contend that university students encounter difficulties with statistics because their knowledge of statistical concepts is not connected. Presumably, these difficulties arise because students are not taught to make these connections, or they are unaccustomed to having to think about the relationships on assessments, or more likely, both.

In this paper, we focus on two assessments that are specifically designed to assess different but related aspects of individuals' knowledge representations: (a) problem sorts which measure how individuals represent a specific aspect of their knowledge in the context of word problems (i.e., representation of inferential statistics problems) and (b) concept maps which measure how individuals represent their knowledge of the discipline as a whole (i.e., representation of the domain of statistics, which includes descriptive and inferential statistics). In this case, the domain is introductory statistics at the undergraduate level. Each type of representation is dependent on how individuals understand relevant concepts and the relationship between each of the concepts, or in Schau and Mattern's (1997) terminology, whether or not they have attained *connected understanding.*

Representation in the Context of Word Problems

The use of word problems to evaluate university students' understanding of introductory statistics is common. Typically, a lesson is given to teach students a particular statistical analysis, such as a *t*-test. Word problems are then assigned to provide practice in performing specific calculations and plugging numbers into formulas for that analysis. Each new analysis is taught separately. Little attempt is made to integrate concepts pertaining to each analysis in the instruction or in assessments. This lack of integration does not provide students with an opportunity to think about all the analyses at one time. Consequently, they do not learn why a problem can be best solved with a particular analysis (e.g., *t*-test). Statistics students rarely learn to make decisions about which analysis is appropriate for a given problem or how a given set of data can be analyzed in more than one way despite the fact that such decision making is at the heart of statistical problem solving (Lovett, 2001; Lovett & Greenhouse, 2000). In this sense, the current use of word problems in statistics classrooms is fairly restrictive—it does little to elicit and provide information about how students represent their knowledge of statistics in realistic problem solving situations. A more effective use of word problems for assessing student representations might be to present students with different types of problems (e.g., *t*-test, chi-square, *f*-test) in several sessions as they learn new statistical analyses, rather than focusing on a specific statistical analysis (e.g., *t*-test) in each lesson. Students could be asked to sort the problems into groups or categories based on similarity. This task, known as the problem (or card) sorting task, puts students in a position in which they must recognize the structural similarities between problems, a necessary step in deciding upon an appropriate

analysis. The sorting task has been used extensively in identifying problem representations in chess, physics, and mathematics (Chase & Simon, 1973; Chi et al., 1981; Quilici & Mayer, 1996; Schoenfeld & Herrmann, 1982; Silver, 1981). Although it has been used primarily for research purposes, this task has value as a classroom assessment in that it can be used to inform both teachers and students of progress in developing problem representations that facilitate problem solving.

How students represent their knowledge when dealing with problems is referred to as a problem representation. A problem representation is a fundamental aspect of problem solving and results from a process of trying to connect one's content knowledge to requirements of the problem before executing a solution procedure (Silver & Marshall, 1990). A problem representation is an example of knowledge in use, that is, how an individual's representation of concepts in a domain applies to a particular problem. Little is known about students' problem representations of statistics concepts. However, research involving other domains makes it clear that the appropriateness of an individual's problem representation is related to expertise (Chase & Simon, 1973; Chi et al., 1981; Larkin, McDermott, Simon, & Simon, 1980; Schoenfeld & Herrmann, 1982; Silver, 1981). In other words, the more accurate an individual's problem representation, the more expert-like his or her problem solving. This knowledge and proficiency develops with experience. Knowledge of concepts and procedures in a domain become increasingly interconnected as competence is achieved (Glaser, 1989). These connections reflect an organized knowledge base that is consistent with the structure of the domain. Many domains (e.g., physics or mathematics) are structured around principles (e.g., Conservation of Energy in physics) or methods (e.g., addition in mathematics). Experts represent or understand problems in terms of principles or methods, which enables them to solve problems successfully (i.e., a principled representation). In contrast, novices often focus on surface features of the problem (i.e., a superficial representation), such as the story line or content irrelevant to solving the problem (e.g., diagrams associated with physics problems, such as an inclined plane), because their knowledge is fragmented and structured in a way that is inconsistent with the principles in the domain. Consequently, novices' success in solving a problem is reduced.

A problem sorting task is an indirect measure of an individual's problem representation (Hassebrock, Johnson, Bullemer, Fox, & Moller, 1993; Ruiz-Primo & Shavelson, 1996). Inferences about how an individual understands a problem

domain are made based on the problems that are grouped together. The assumption is that if all problems designed to represent a statistical analysis (e.g., *t*-test) are grouped together, then the sort was performed on a principled basis by thinking of relevant content. If, on the other hand, problems representing common features of a particular story line are grouped together, the assumption is that the sort was not related to content and performed on a superficial basis by focusing on cues in the problem. More direct measures are needed to ensure the validity of these assumptions. Student explanations of their sorts can help identify the specific reasons for the groupings (Lavigne & Glaser, 2001; Quilici & Mayer, 1996). However, a problem representation reflects only a small part of an individual's knowledge base since it focuses on a subset of concepts in the domain and on specific relationships between these concepts. An additional measure is needed to provide a broader and more explicit picture of how students represent their knowledge of introductory statistics. One assessment task that can provide in-depth and direct information about students' representation of concepts in a domain is the concept map (Novak, 1995; Ruiz-Primo & Shavelson, 1996; Schau, Mattern, Zeilik, Teague, & Weber, 2001).

Representation in the Context of Concept Maps

A concept map is used to represent and organize knowledge; it is a visual representation of how a student understands concepts and their relationships (Novak, 1995). Concepts are represented in the form of nodes and relationships in the form of links. A label on a link identifies the nature of the relationship between the concepts it connects. Cross-links depict relationships between concepts in different parts of the concept map. Specific examples of concepts can also be included in a concept map. Concept maps have one of the following two structures depending on the theoretical perspective and on how the domain itself is organized (Kinchin, Hay, & Adams, 2000): (a) a hierarchical structure in which concepts are subsumed under superordinate concepts or (b) a nonhierarchical structure in which concepts are surrounded by other concepts or depicted in a series. Note that some researchers use "knowledge map" as a more general term to indicate that conceptual knowledge is not the only type that can be represented (Klein, Chung, Osmundson, Herl, & O'Neil, 2002). Another distinction between concept and knowledge maps is that a standard set of labeled links is used across maps in the latter but not the former (O'Donnell, Dansereau, & Hall, 2002). We prefer to use the more common

term *concept map* to avoid the confusion that may arise with the different usage of the term *knowledge map*.

A concept map is conceived as representing an individual's cognitive structure, that is, a framework in which concepts are organized in memory (Ruiz-Primo & Shavelson, 1996). Ausubel assumed that this organization was hierarchical and that students learn content by (a) making connections between new information and general concepts such that the former is subsumed under the latter (i.e., subsumption), (b) distinguishing the meaning of different concepts as new relationships are formed (i.e., progressive differentiation), and (c) recognizing new relationships between groups of three or more concepts (i.e., integrative reconciliation) (Rafferty & Fleschner, 1993; Ruiz-Primo & Shavelson, 1996). This perspective underlies hierarchical concept maps. Nonhierarchical concept maps are more consistent with the associationist viewpoint, which does not assume cognitive structures to be hierarchical (Ruiz-Primo & Shavelson, 1996). Note that the manner in which information is represented is dependent on the content that individuals interact with. For example, statistics is a hierarchical domain that lends itself to hierarchical organization whereas the "water cycle" is more nonhierarchical in nature and thus best represented in that manner (Ruiz-Primo, Schultz, Li, & Shavelson, 2001). Finally, the Ausubel and associationist theories are consistent with schema theory in that they posit a cognitive structure in which concepts are organized in memory, and they focus on the connections learners make between concepts and the process by which meanings are generated and revised.

Indeed research in this area suggests that concept maps can result in meaningful learning when used as advance organizers for content to be learned and when created by students to illustrate their knowledge of concepts in a particular domain. In the case of viewing maps, O'Donnell et al. (2002) found that knowledge maps, in which standard labeled links are used, are more effective than text in facilitating recall of main ideas, particularly for students whose verbal skills or prior knowledge is weak. Wells (1999) found that advance organizers significantly enhanced college students' conceptual understanding of biology. When required to construct concept maps, elementary students' understanding of relationships between biology concepts is enhanced (Stice & Alvarez, 1987) and secondary students' anxiety about biology is reduced, which in turn, enhances their learning of biology (Jegede, Alaiyemola, & Okebukola, 1990).

As assessment tools, concept maps can differentiate ability levels (Klein et al., 2002) and can be a valid indicator of students' representation of their knowledge (Schau & Mattern, 1997), particularly when coupled with written essays in mathematics courses (Bolte, 1999). Concept maps assess representations more explicitly than essays and multiple-choice tests because they are more commonly employed to elicit conceptual understanding (Klein et al., 2002) and reveal student misconceptions that can be masked by traditional measures (Roberts, 1999). Moreover, concept maps have face validity because they can facilitate learning by linking instruction and assessment. Nonetheless, a tremendous amount of work is needed to assess the validity and reliability of concept maps, which vary in format and scoring systems (McClure, Sonak, & Suen, 1999; Rice, Ryan, & Samson, 1998; Ruiz-Primo & Shavelson, 1996). Moreover, students need training in constructing concept maps prior to engaging in the task—at least 45 minutes (Ruiz-Primo et al., 2001; Stoddart, Abrams, Gasper, & Canaday, 2000). In comparing two formats, fill-in-the-map (i.e., a pre-established structure with some concepts or links identified is presented to students who have to fill in the rest of the map from a list of concepts or links) and construct-a-map (i.e., no information about structure, concepts, or links provided), Ruiz-Primo et al. (2001) found that the latter is a more valid indicator of students' representations because it reflects differences in the structure of students' knowledge. Finally, McClure et al. (1999) examined a variety of scoring methods and found that the most reliable technique involved using a master map. Expert maps are often employed to assess students' representations (Stoddart et al., 2000) since the assumption is that the structure of students' concept maps will become more consistent with those of experts as their knowledge increases (Ruiz-Primo et al., 2001).

Case Study Aims

This case study is meant to be exploratory. We were especially interested in how a statistics expert would perform on the problem sorting and concept mapping tasks. For example, would the inferential statistics problems be sorted in the manner expected based on how the problems were designed? (e.g., all the F-test problems sorted together). Would the expert concentrate on the same problem features as did the designers of the problems? In addition, would the expert's representation on the problem sorts be consistent with his representation of statistics as depicted in the concept maps? Finally, because we were also interested in obtaining data that would enable us to develop scoring rubrics for assessing student representations of

statistics, we wanted to know the different ways in which the expert expected introductory statistics students to perform on the problem sorting task.

Method

Participants

Three content area experts were contacted to perform the two representation tasks. All experts were faculty members at different universities and departments (i.e., Departments of Statistics, Educational Psychology, and Educational and Counselling Psychology). Selection of the experts was based on their statistical experience and willingness to participate. This was thus a convenience sample. Although all three experts agreed to participate in the study, only one performed the tasks at the time of this report. This participant was a male professor, age 44, who was trained as a statistician and had approximately 14 years of experience in teaching statistics (4 years at the undergraduate level and 10 years at the graduate level). He has taught introductory statistics to undergraduate students in psychology and engineering (e.g., introduction to statistics I & II, engineering / business statistics, and engineering probability), as well as a number of advanced statistics courses to undergraduate and graduate students (e.g., linear regression, project courses, and graduate theory and methodology). His approach to teaching statistics is both mathematical and applied. He explains this approach in the following way: "The care and self-monitoring of mathematical thinking are important in statistics. However, statistics' raison d'être is applications, and this is also an effective way to capture students' intellectual attention for the subject."

Materials

Statistics problems. Twelve statistics word problems were typed and presented to the participant on separate 3 x 5 index cards. Problems represented four inferential tests that are commonly taught in an introductory statistics class, namely, *t*-test, *f*-test, chi-square test, and correlation (see Appendix A). Three problems were developed for each test, resulting in a total of 12 problems (adapted from Lovett, 2001; Quilici & Mayer, 1996). These problems can be represented in terms of two overarching features: structural and surface. Structural features represent fundamental statistical analyses. Surface features represent a superficial understanding of statistics and include a focus on semantic (e.g., topic or cover story) or literal (e.g., data organization) similarities. Order of presentation was

determined by randomly selecting problems from a hat. The resulting sequence was presented to each participant in the same order.

Concept mapping training. The participant was given a 45- to 50-minute training package for constructing concept maps prior to performing the concept mapping task. The training materials were adapted from Ruiz-Primo, Schultz, and Shavelson (2001) and involved the following: (a) introducing participants to concept mapping by describing the purpose, components (i.e., nodes, links, propositions, cross-links, and examples), and structures of a concept map (i.e., hierarchical and nonhierarchical) as well as providing examples; and (b) providing an opportunity to construct concept maps by identifying relationships between pairs of concepts (e.g., earth, solar system), constructing propositions, redrawing a concept map, and constructing two maps based on two different lists of concepts (i.e., human body and teaching). The practice activities involved content familiar to most individuals but was not related to the domain (i.e., statistics) to be represented on the concept mapping task.

Measures. Four measures were used in this study: a background questionnaire, a sheet for documenting problem sorts, concept maps, and structured interview questions. A background questionnaire was administered to identify demographic information such as the participant's age, gender, position at the university (i.e., professor, associate professor, and assistant professor), and department. Information pertaining to the participant's experience was also solicited, specifically, the total number of years in teaching statistics, the statistics courses taught, the level at which the courses were taught (i.e., undergraduate and/or graduate), and the number of years teaching each course. Finally, the questionnaire required that the participant select one or more of the following approaches that he used in teaching statistics: mathematical, application, and theoretical. Space was provided for making comments.

The second measure, a sheet for documenting problem sorts, was designed to assess the accuracy of the participant's representation of inferential statistics problems and to generate a range of expected student performance on the same task. The participant was required to sort the problems in his own way first, then to identify the sorts he would expect students taking introductory statistics to produce, and finally, to rank these alternative sorts in order of relevance (i.e., to the task) or level of sophistication. The documentation sheet provided tables for indicating the problem numbers in each pile for the first two parts of the task and a space was

provided for writing the rank. The third measure, a concept map, was designed to assess the organization or structure of the participant's knowledge of introductory statistics as a whole and to give a sense of the participant's domain knowledge.

The fourth measure, structured interview questions, was designed to assess the nature of the participant's representations. Questions were displayed on the instruction sheet for each task. The following structured interview questions pertained to the problem sorting task: (a) Why is this the best way to sort these problems? (b) Why did you sort the problems into these particular groups or piles (please say aloud which problems you are referring to)? How are the problems in each pile similar? (c) How does each group of problems or sort differ from each other (again please indicate which set of problems you are referring to in your explanation)? and (d) Which problems did you have difficulty with (if any) and why? In other words, were any of the problems ambiguous or problematic? The structured interview questions for the concept mapping tasks included the following: (a) Why did you decide on this organization or structure for the concept map? (b) Why are these concepts the most important? Were there any other concepts that you were unsure about including in the concept map? Why? and (c) What do you think are the most critical relationships in this concept map?

Procedure

The participant was mailed the materials (i.e., instructions, a background questionnaire, 12 index cards, a training package, a sheet for documenting problem sorts, and an audiotape) after consenting to participate in the pilot study. Instructions were provided for each task. First, the participant was instructed to fill out the background questionnaire. Second, he was required to use the deck of index cards (each numbered in the presentation sequence) to sort the problems based on how they "best went together" (Mariné & Enscribe, 1994; Quilici & Mayer, 1996). The participant was not required to actually solve the problems but to categorize them in the most important way(s) that he thought they were similar. A documentation sheet was provided for reporting the problem sorts. Structured interview questions and an audiotape were provided for obtaining and recording the participant's explanations of his problem sorts. Third, the participant was instructed to identify the number of different ways he expected introductory statistics students to sort the same problems. He was then asked to rank order these student groupings in terms of relevance to the task or level of sophistication. The possible student sorts and their ranking were reported on the documentation sheet

and the participant's responses to structured interview questions regarding these alternative sorts recorded on audiotape.

Fourth, the participant was given some experience in concept mapping prior to constructing his own map of statistics in the form of a 45- to 50-minute training package. He was instructed to commence the concept map task once he felt comfortable with the process. Concept maps could be drawn using paper and pencil or a computer. The first concept mapping task involved constructing a map that identifies the main concepts in introductory statistics (descriptive and inferential) and shows how each of the concepts are related to one another. The importance of drawing a map that reflects how one thinks the domain is organized was emphasized. The participant was instructed to turn on the tape recorder and to respond to the structured interview questions upon completion of the concept map. The participant was also required to draw a second map, but this time focusing on inferential statistics only. He was informed to bypass this second task if the first concept map he constructed included *all* the concepts and relationships that he considered critical to inferential statistics. However, if the participant did not go into any detail regarding inferential statistics on the first map, then he was asked to produce a second concept map that did. Upon completion of this concept map, the participant was once again required to turn on the tape recorder and to respond to the same questions that were presented in the first concept mapping task.

The participant began with the problem sorting tasks and finished with the concept mapping tasks. Task order was not counterbalanced because such differences were not expected with experts. More specifically, the concept mapping task was not expected to influence the expert's performance on the problem sorting task, since experts' knowledge is sufficiently structured and integrated so that an advanced organizer is unnecessary for successful performance. In contrast, such cues would be expected to result in task differences for students who, as novices, would benefit from starting with a concept mapping task where the resulting map could be used as a reference point for performing the problem sorting task.

Results

The audio data were transcribed and segmented according to the participant's explanation for each sort or feature of the concept maps. In the case of problem sorts, the verbal data was segmented based on the explanation for each problem sort or pile of problems, regardless of the number of problems grouped together in each

pile. In other words, five sorts yielded at least five explanations or segments of data. Note that we use the qualifier "at least" here to indicate that additional explanations are possible, particularly if the participant changes his mind about where certain problems belong and/or if he summarizes his explanations at the end. In the case of concept maps, the verbal explanations were segmented based on how the participant compartmentalized his description and explanation of the concept maps. Since the participant divided his explanations of the maps by the number of columns appearing in the representations, the verbal data were segmented according to these columns. In other words, five columns yielded at least five segments of verbal data. Given that this is a case study of one participant, the data were examined qualitatively. Consequently, the results are presented descriptively. We first present the results pertaining to the problem sorts followed by the concept maps produced by the participant.

Problem Sorts

In this section, we describe the participant's performance on the problem sorting task along with his explanations for the sorts. We then describe the various levels of sorting performance he expected of undergraduate students taking introductory statistics.

Expert performance. The inferential statistics problems were manipulated to be categorized in a certain way, namely, by type of statistical analysis. That is, all problems requiring an F-test were intended to go together, all chi-square problems were designed to go together, etc. Table 1 demonstrates that the participant did not sort the problems in the intended manner. One difference between the intended and expert sorts is the total number of groupings produced and the number of problems comprising each sort. Problems were designed to be sorted in four ways, resulting in

Table 1

Intended Problem Sorts vs. Expert Sorts

# of sorts	Intended sorts	Initial expert sorts	Final expert sorts
Sort 1	2, 5, 7	**2, 5**, 9	2
Sort 2	1, 10, 12	6, 12	6, 12
Sort 3	3, 4, 9	4	4
Sort 4	6, 8, 11	3, 8, 10	3, 5, 8, 9, 10
Sort 5		1, 7, 11	1, 7, 11

four different groups consisting of three problems each. In contrast, the participant sorted problems into five piles with a different number of problems in each sort. Note that the number of problems in each sort was closer to the expected three in the participant's initial grouping (i.e., initial expert sort column) than in his final sort (i.e., final expert sorts column) in which the number of problems in each sort varied from one to five as a result of removing two problems from one sort (i.e., sort 1) and adding them to another already comprising of three problems (i.e., sort 4). The second difference is that the actual problems designed to belong to specific categories were not sorted in the same way by the expert. Initially, only two of the 12 problems (17%, problems 2 and 5 in sort 1) designed to be in the same category were sorted together. However, none of the problems were grouped together as expected in the final sort (0%).

The source of these differences lies in the rationale for why the problems were designed to be in the same category and in the participant's explanation of his sorts. The problems were designed to vary by type of statistical analysis required to solve the problems. Four types of analyses were represented in the problems: *t*-test, chi-square, *f*-test, and correlation. Possible features of the problems that may facilitate deciding which analysis is appropriate for a particular problem are the purpose of analysis (i.e., to examine a difference or a relationship), data type (i.e., measurement or categorical), and number of groups or variables (i.e., two or multiple). Table 2 displays how these features are associated with each type of statistical analysis (see the second column labeled "Intended sorts"). The expectation was that experts would sort the problems based on statistical analysis. However, the participant's verbal explanation for his sorts indicated that he was primarily concerned with the kinds of threats that the problems presented to the validity of inference that could be made for each analysis. The type of analysis that would be appropriate for each problem was a secondary concern. His approach to the task was to first sort the problems based on whether or not the problems involved experimental or undermine a regression analysis of each of the variables… Time series is data that is observational data. Then he sorted the observational problems (he identified only one problem as being experimental) based on where the data came from and where he expected threats to validity of inference.

Table 2

Intended vs. Expert Reasons for Categorizing Problems

# of sorts	Intended sorts	Initial expert sorts	Final expert sorts
Sort 1	*t*-test: difference / 2 groups / measurement data	Observational data / *f*-test (ANOVA)	*f*-test (ANOVA) / observational data
Sort 2	Chi-square: relationship / 2 variables / categorical data	Observational data / Spatial data (potential problem in analysis) (different analyses)	Observational data / Spatial data (potential problem in analysis) (different analyses)
Sort 3	*f*-test: difference / multiple groups / measurement data	Experimental data / F-test (one-way ANOVA)	Experimental data / F-test (one-way ANOVA)
Sort 4	Correlation: relationship / 2 variables / measurement data	Observational data / Time series (details of analysis differ)	Observational data / Time series (details of analysis differ)
Sort 5		Observational data / Survey data (different analyses)	Observational data / Survey data (different analyses)

As shown in Table 2, problems involving observational data were grouped according to three threats to validity that could possibly undermine a statistical analysis: spatial dependence between observations (i.e., sort 2), time dependence between variables (i.e., sort 4), and threats to validity (e.g., convenience sample) specifically related to survey sampling (i.e., sort 5). By spatial dependence, the participant meant the following:

> "The idea in each case is that the observation in one location might be somehow related to observations in another location and this complicates a careful analysis. For example, New York and New Jersey [problem 6] might be more similar than you would expect if they were actually independent observations or something like that. Ah, and certainly more similar than New Jersey and Kansas. This kind of spatial dependence between observations that are close together spatially would potentially create a problem in the analysis."

Time dependence was explained as the following: "In problem 8 you have a dependence from one year to the next. So there's time series dependence that can collected over time." Finally, the main threat to inferences made based on survey sampling was use of a convenience sample, which can limit generalization.

It is especially noteworthy that the participant did consider the type of analysis that would be required for each problem. However, his overarching criterion consisted of the threats that might undermine the validity of the inferences made from the analyses. The participant was quite explicit about the fact that the problems in these sorts required different analyses and at times even identified the type of statistic required. When problems are regrouped according to the type of statistical analysis suggested by the participant, and threats to validity and the observational/experimental distinction are disregarded, we see that the participant's performance was more consistent to the expected performance based on the design of the problems than when only the problem sorts were examined (i.e., without the explanations).

Table 3 presents a regrouping of the problems based on the participant's identification of appropriate statistical analyses, which may be undermined by specific validity threats. Unfortunately, the participant did not identify an analysis for problem 3, merely indicating that it involves a comparison. He also indicated that problem 10 reflected an interest in "outcome" a term usually identified with chi-square and thus labeled as such. According to this table, five of the twelve problems were correctly identified as requiring the same analysis (42%), assuming of course, that there is nothing misleading about the problems themselves. It is important to note that many statisticians now use ANOVA or F-test instead of a t-test. Moreover, it seems that the participant may have subsumed the t-test within ANOVA models, which he identifies as a main form of analysis in his last concept map (as will be seen later). Taking this possibility into account, the consistency between the participant and designed sorts increases to eight of twelve problems (67%). Finally, regression and correlation are not that different in that they are in the same family of analyses, but serve different functions. It is interesting that the participant identified

Table 3

Regrouping of Problems Based on Type of Statistical Analysis Only

# of Sorts	Intended Sort	Final Expert Sort
Sort 1	*t*-test: 2, 5, 7	Logistic regression: 7
Sort 2	Chi-square: 1, 10, 12	Chi-square: **1, 10, 12**
Sort 3	*f*-test: 3, 4, 9	*f*-test: 2, **4**, 5, **9**
Sort 4	Correlation: 6, 8, 11	Regression: **6, 8, 11**

all correlation problems as requiring a regression analysis. His representation of inferential statistics in terms of analysis selection suggests that he may have viewed correlation as belonging to the category "regression models." If this is indeed the case, then the participant was 92% consistent with the expected or ideal sort (taking into consideration that t-test and ANOVA may not be distinguished).

Expected student performance. The participant produced three sorts that he thought students might produce and ranked them from least sophisticated to most sophisticated. The least sophisticated method he expected students to group the problems was to sort based on superficial features. Specifically, he expected problems to be sorted into four groups, each having to do with one the following features: professor, government, weather, and survey. The participant explained that this kind of sort is good in the sense that "attention is paid to the theory of application of the problem although even that's pretty superficial here since the difference between the government problems and the survey problems maybe isn't that great in terms of application." He felt application was important in data analysis. Nonetheless, the sort was seen as superficial and not particularly sophisticated because it did not provide any guidance about how to analyze problems and do the statistics.

A sort at a higher level of sophistication consisted of producing two piles: one for experimental data (one problem) and the second for observational data (11 problems). The participant thought that this sort was more sophisticated than the previous one because thinking about data in this way cues to possible threats to the validity of the analysis. However, it is not the most sophisticated because it is not informative of the kind of analysis that is needed. The most sophisticated sort the participant expected students to produce involved identifying the independent and dependent variables and deciding whether they are discrete (i.e., categorical) or continuous (i.e., measurement). He indicated that he and his colleagues hope that their students can think about these features after a semester of statistics. He explained that this kind of approach to sorting problems was useful because it "immediately suggests both the EDA [Exploratory Data Analysis] techniques [such as graphs] and more formal analytic techniques that can be used to analyze data." For example, if the independent variable is discrete and the dependent variable is continuous then the graph to use is a box plot and the analysis to perform is analysis of variance. The participant produced four sorts of problems reflecting the combinations of discrete and continuous variables: (a) continuous

independent/continuous dependent, (b) discrete independent/continuous dependent, (c) continuous independent/discrete dependent, and (d) discrete independent/discrete dependent. A fifth pile consisting of one problem was produced because the participant felt that it did not fit well with the above categories.

At the end of the task, however, the participant suggested a fourth sort that combined features of the second and third proposed sorts. This fourth sort involves categorizing the problems based on whether they involve categorical or continuous data for each of the variables, which suggests the graphical and analytic techniques, and then deciding whether the data are experimental or observational, which suggests various threats to the validity of the analysis. According to the participant, these two sets of features provide the most information about how to proceed with any given problem. It can therefore be classified as the most sophisticated. Table 4 summarizes these four approaches to sorting statistics problems.

Concept Maps

The participant constructed three concept maps. The first map represented his understanding of introductory statistics and displayed the general concepts in the domain (see Figure 1). The second concept map elaborated on inferential statistics (see Figure 2) and the third map elaborated on one aspect of inferential statistics, namely, types of statistical analysis (see Figure 3). These last two concept maps are particularly important for understanding the participant's representation of inferential statistics on the problem sorting task. Although the participant was given instruction on how to construct concept maps, he felt he could best represent statistics by constructing a horizontal organization with columns of concepts that are equivalent (i.e., occur at similar stages of problem solving). He called the maps "a cross between a concept map and a flow chart" and indicates how one should proceed in solving statistical problems. Figure 1 illustrates the participant's overall or general concept map.

The participant organized the representation around five sets of concepts, which were organized and discussed in terms of columns. The structure of the representation can be thought of as hierarchical if the first set of concepts shown in the first column (i.e., problem and data) are treated as the broad concepts and concepts in the second column (i.e., background story, questions, variables, and data collection design) are subsumed under the two general concepts. The columns in the

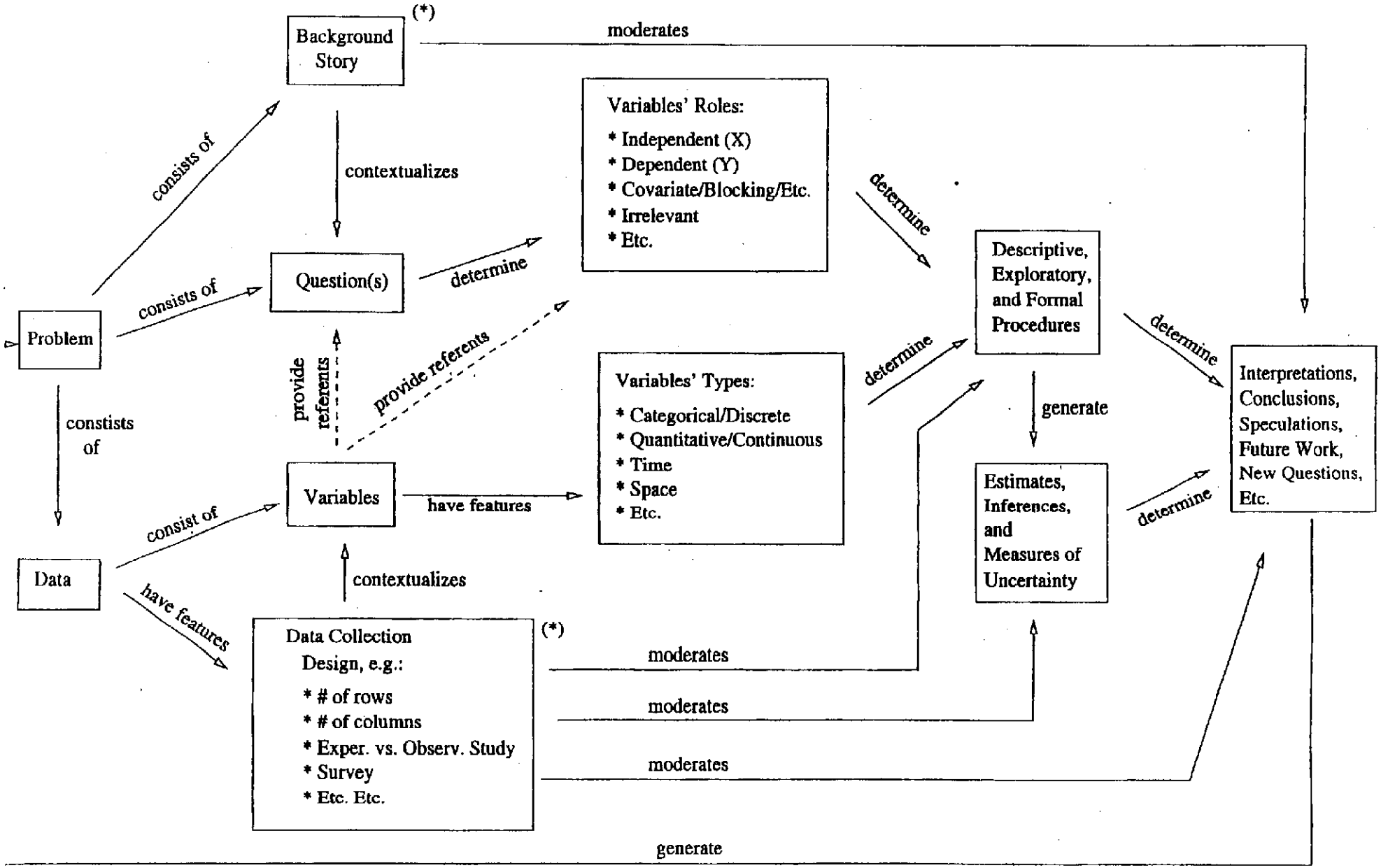
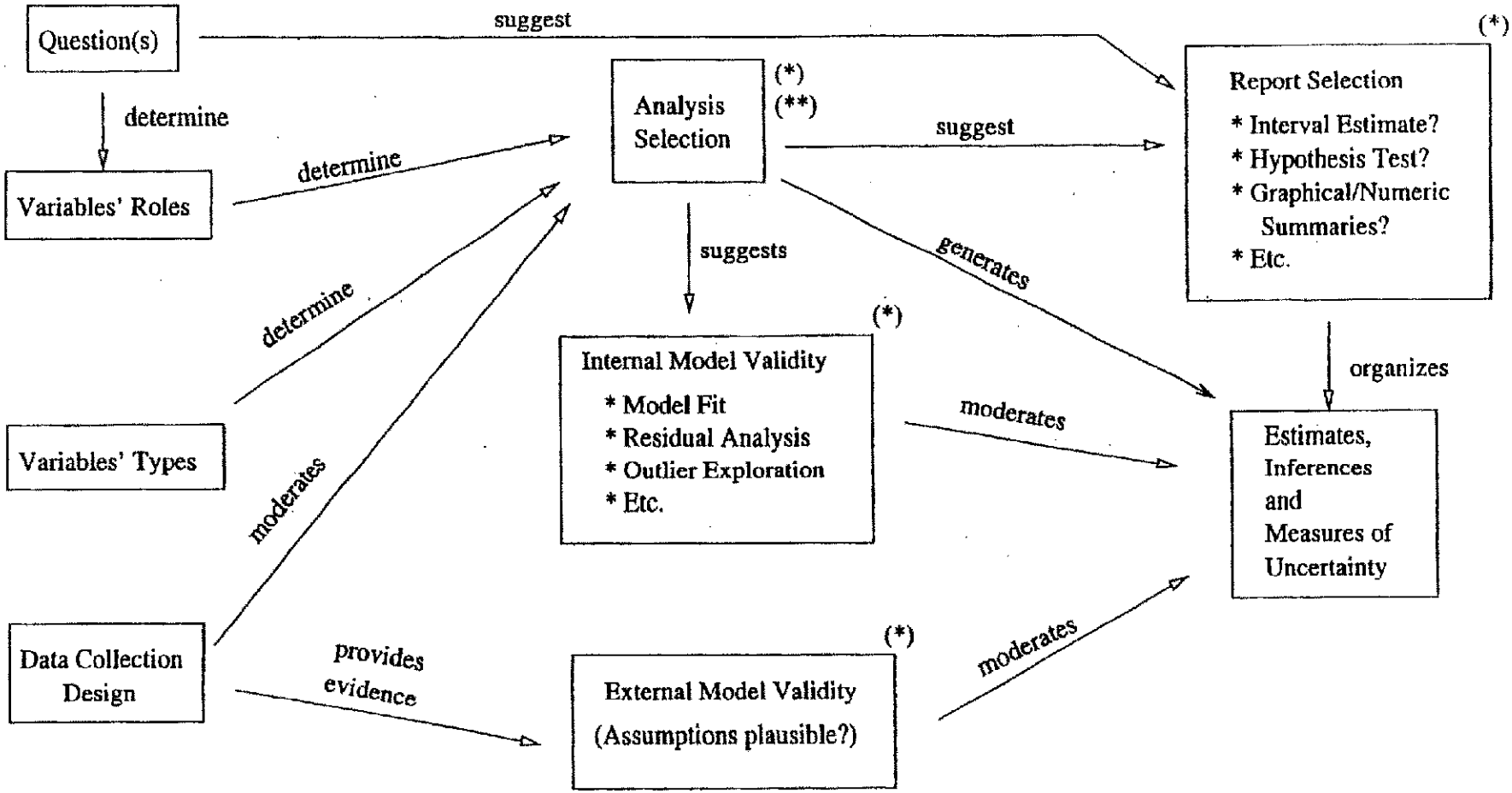


Figure 1. Participant's overall concept map.



(*) These boxes elaborate "Descriptive, Exploratory and Forward Procedures.

(**) See next page for a partial elaboration of "Analysis Selection"

Figure 2. Participant's concept map for inferential statistics.

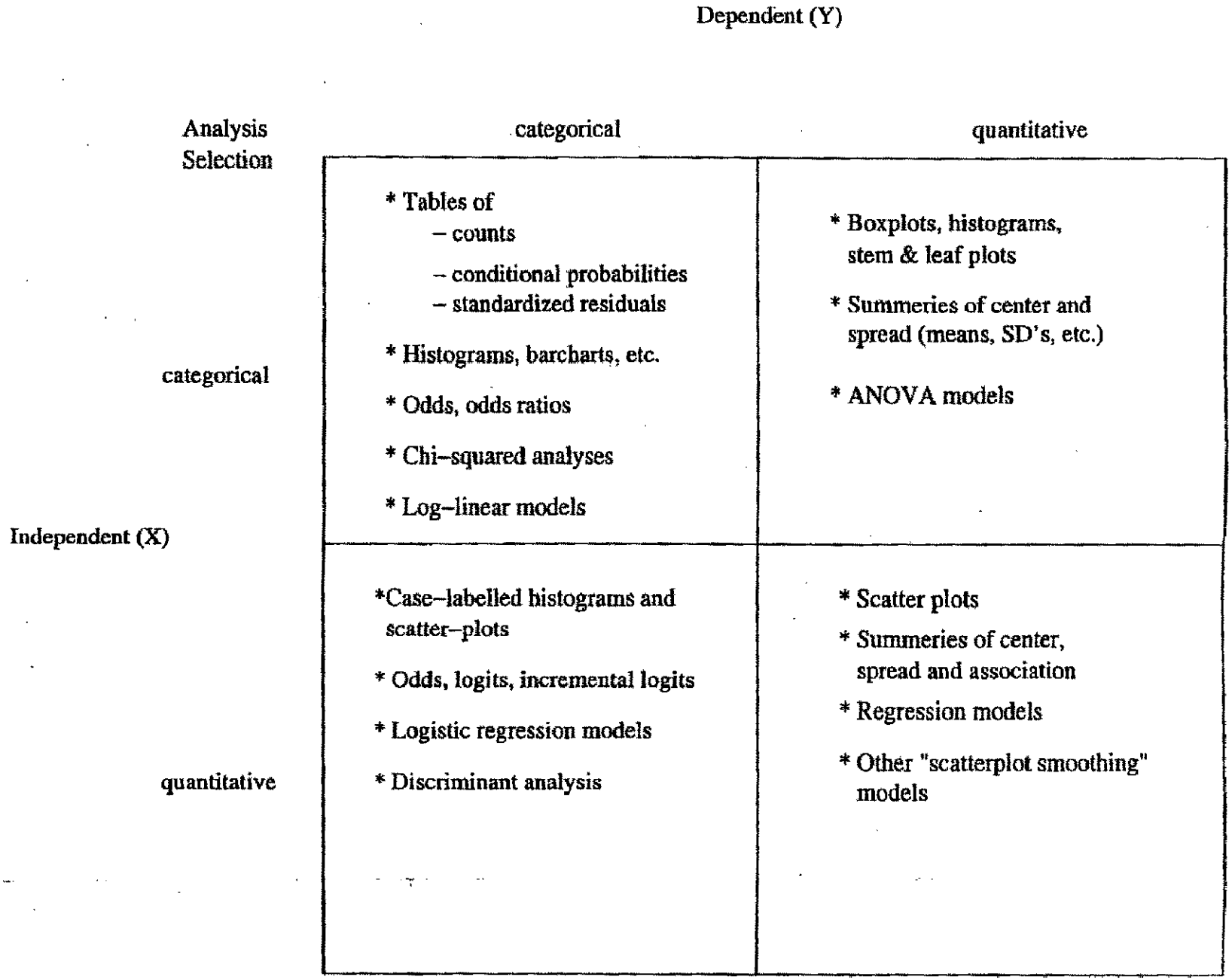


Figure 3. Participant's concept map for statistical analysis.

Table 4

Problem Sorts Expected of Students at Various Level of Sophistication

Level of sophistication	Reason for categorization
Least sophisticated: 1	Superficial features: cover story
Relatively sophisticated: 2	Identifying whether data is experimental or observational
Highly sophisticated: 3	Identifying the variable (i.e., independent, dependent) and deciding which is discrete (i.e., categorical) and which is continuous (i.e., measurement)
Most sophisticated: 4	Identifying whether data is experimental or observational *and* Identifying the variable (i.e., independent, dependent) and deciding which is discrete (i.e., categorical) and which is continuous (i.e., measurement)

representation can be thought of as levels within the hierarchy. The fact that the participant chose to represent the domain partly as a flowchart and that it began with the concepts "problem" and "data" illustrate a research methods or problem solving approach to statistics. According to the participant, problem and data are the basic concepts in statistics and starting point in statistical problem solving. One begins with a verbal description and some information about the data that could be collected. These concepts are related to a background story, question(s), variables, and data collection design. The variables provide referents to the variable roles and have the feature variables types. Notice that the participant's categorization of problems on the previous task focused on the data collection design and variable types. Specifically, he emphasized whether the design was for an experimental or an observational study and whether the variables were constrained by time series or spatial dependencies. This constraint is represented by the "moderates" relationship depicted in the arrows going from the data collection design to the remaining concepts on the right hand side of the representation. The participant explains the relationship "moderates" as the following:

> "What I mean by moderate is that by referring back to the data collection design one either has greater or lesser confidence in one's selection of a procedure and greater or lesser confidence or somehow one gains some qualification about the estimates, inferences, and measures of uncertainty that were generated from the analysis, from the procedures."

Another interesting feature of the representation is that the interpretations, conclusions, speculations etc. need to be translated back into the nonstatistical language of the problem (e.g., science, educational). As such, there is a link back to "background story" and to the concept "problem" to represent the cycle of discovery. And so the cycle continues. According to the participant, he would be pleased if first year students could understand statistics in the manner represented in Figure 1.

Figure 2 illustrates the participant's elaboration of the general concept map by concentrating on inferential statistics. Two comments are needed at this point to explain how this second concept map relates to the first representation. First, the concepts in the first column of the inferential statistics representation combine concepts depicted in the second and third columns of the general concept map. That is, the concepts "variable roles" and "variable types" in the third column of the general map are actually an expansion of the concept "variable" which appears in the second column of that representation. In the subsequent representation, which focuses on inferential statistics, the participant moved the two expanded variable concepts into the same column (i.e., in the first column of the inferential statistics map which is the equivalent of the second column in the general statistics representation). Second, the concepts shown in the second column of the inferential statistics map actually unpack or elaborate on the concept "descriptive, exploratory, and formal procedures," which is depicted in the fourth column of the general map.

According to the participant, five major concepts comprise the essence of inferential statistics: analysis selection, internal model validity, external model validity, report selection, and estimates, inferences, and measures of uncertainty. These are of course, related to four overarching concepts that he felt are involved in all aspects of statistics, that is, question(s), variables' roles, variables' types, and data collection design. Two aspects of this representation are interesting. First, the participant's constant concern with validity is evident in his explicit representation of two additional validity concepts, namely, internal model validity and external model validity. Second, the participant's view of statistics as a problem solving approach is again reflected in the concept "report selection," which does not appear to be a common explicit consideration in most statistics courses. The participant concludes his description of the inferential statistics representation by saying that he "would just be thrilled if statistics students actually understood that this was what was underlying a lot of what we do." He did not want to focus on the technical

nature of inferential statistics because his view of statistics, as shown in his representations, is conceptual and this is the learning outcome he hopes his students can attain.

In his third representation (see Figure 3), the participant elaborated on the "analysis selection" concept shown in the second column of the inferential statistics map. Surprisingly, he chose to represent this aspect of statistics as a two by two table. Again, the participant is interested in the "big picture" rather than in the technical details. He felt that students too often get lost in the detail because the technical aspects of the discipline dominate the concepts when they should be secondary to the big picture. This representation also illustrates the concepts that need to be considered in making decisions about the types of analyses that may be appropriate for given problems. This representation is meant to show how one goes from the concepts to the procedures, hence the contingency table representing analysis decisions. Notice that the main concepts involved in this type of decision making consist of variable types (i.e., categorical / discrete and quantitative / continuous), variable roles (i.e., independent X and dependent Y), and analysis selection. The analyses identified are both formal (e.g., chi-square) and informal (e.g., histogram). Surprisingly, correlation was not identified as a possible analysis for the quantitative / quantitative quadrant of the representation, which suggests that the participant may have grouped this type of analysis under "regression models" or "other scatterplot smoothing models."

Discussion

The current assessment challenge is to build assessments based on empirical and theoretical knowledge of learning and cognition. A cognitive proficiency that has received relatively little attention in statistics classrooms is the nature of students' representations of problems and representations of content. A problem representation is an instantiation of an individual's representation of content in that it reflects a particular kind of knowledge in use, usually a specific set of concepts in the discipline. A small number of researchers are exploring students' representations of statistics problems (e.g., Lavigne & Glaser, 2001; Quilici & Mayer, 1996; Quilici & Mayer, 2002) and assessing the validity of concept maps as measures of representation or connected understanding (e.g., Schau & Mattern, 1997; Schau et al., 2001). However, instructors rarely provide learners with opportunities to perform these representation tasks in the classroom.

In this paper, we focused on two assessments that are specifically designed to assess different but related aspects of an expert's knowledge representation in statistics: (a) problem sorts which measure how individuals represent a specific aspect of their knowledge in the context of word problems (i.e., representation of inferential statistics problems) and (b) concept maps which measure how individuals represent their knowledge of the discipline as a whole (i.e., representation of the domain of statistics, which includes descriptive and inferential statistics). We were especially interested in exploring the following questions: (a) would the inferential statistics problems be sorted in the expected manner based on how the problems were designed (e.g., all the F-test problems sorted together)? (b) would the expert concentrate on the same problem features as did the designers of the problems? (c) would the expert's representation on the problem sorts be consistent with his representation of statistics as depicted in the concept maps? and (d) what would different levels of expected student performance on the problem sorting task look like?

In response to the first two questions, the data suggest that the problem sorting task was a useful measure of representation when it was supplemented with an individual's explanations for his or her sorts. Simply focusing on whether problems designed to be in the same category were sorted together was misleading and underestimated the expert's representation. For example, the expert did not sort the problems in a way that was directly related to the problem solution, that is, type of statistical analysis. However, the expert's explanations for his sorts revealed that he focused on different problem features than had the problem designers. The expert sorted on the basis of the data collection design, specifically whether the study was designed to be experimental or observational, and on potential threats to the validity of inference from an analysis, namely, time series or spatial dependency. These threats were seen as primary and the analyses secondary. In contrast, the problems were designed such that the analyses were primary, with particular attention paid to the following features: purpose of analysis (difference or relationship), data type (categorical or measurement), and number of variables or groups (two or multiple). Thus, the designers and expert focused on completely different features, but all were relevant to solving the problems and were highly sophisticated. Moreover, the expert did consider which type of analysis would be appropriate for most problems in each group and when these responses were taken into account, the expert was found to select the appropriate analyses for most of the problems. This observation

suggests that explanations must accompany problem sorts in order to obtain a valid assessment of problem representations on the task and is consistent with previous research (Lavigne & Glaser, 2001).

The concept map measure was included in the study in order to more directly assess the expert's representation of statistics as a whole and of inferential statistics more specifically, which was the content addressed on the problem sorting task. The expert's concept maps were remarkably consistent with his representation of inferential statistics on the problem sorting task. However, the manner in which this knowledge was structured was only revealed by the concept map and provided a "big picture" of how he conceptualized the domain of statistics. In this sense, the concept map measure can assist in the interpretation of performance on the problem sorting task and provides a comprehensive picture of an individual's knowledge representation and how it is applied on problems. A concern with threats to validity was evident on both the problem sorting and concept mapping tasks. However, this concern was more pronounced on the problem sorting task as it guided performance, whereas concepts related to this threat (i.e., spatial and time dependency) were subsumed under the "variable type" concept in the concept map, its' prominence lessened. It therefore appears that use of both measures can be mutually informative, with the concept map providing a broader picture and the problem sorts illustrating how certain concepts become salient when applied to different contexts.

Finally, asking the expert to estimate how introductory students would sort the problems and to classify the expected responses in terms of level of sophistication is particularly informative for developing scoring rubrics to assess student representations on the same task. The expert produced four levels of performance he deemed acceptable ranging from least sophisticated to most sophisticated. One of these levels (i.e., relatively sophisticated) reflected one feature the expert focused on in his sort, namely, on whether the study had an experimental or observational design. Sorting in this manner was considered somewhat sophisticated because it paid attention to possible threats to validity of inference. Interestingly, the expert explained that a sort that was also based on the variable role (i.e., independent and dependent) and type (categorical and measurement) was important for determining the type of analysis and thus reflected a sophisticated sort. However, he himself never referred to these concepts in explaining which analyses could be performed for each problem. He simply indicated the analysis without articulating his

reasoning why that particular analysis was appropriate. This observation is in line with the expertise literature demonstrating that experts' tacit knowledge is often not articulated.

In conclusion, this case study suggests that multiple measures should seriously be considered when the goal is to obtain a valid assessment of individual's representation. Additional data gathered from multiple experts can certainly facilitate and validate the content of the assessment as well the development of scoring rubrics for examining representations.

References

Bolte, L. A. (1999). Using concept maps and interpretative essays for assessment in mathematics. *School Science and Mathematics, 99* (1), 19-30.

Chase, W. G., & Simon, H. (1973). Perception in chess. *Cognitive Psychology, 4* (1), 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* (2), 121-152.

Garfield, J. (2002). Recent Ph.D.s in statistics education. *Statistics Education Research Journal, 1* (1), 15-16.

Glaser, R. (1990). Toward new models for assessment. *International Journal of Educational Research, 14* (5), 475-483.

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

Hassebrock, F., Johnson, P. E., Bullemer, P., Fox, P. W., & Moller, J. H. (1993). When less is more: Representation and selective memory in expert problem solving. *American Journal of Psychology, 106* (2), 155-189.

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

Jegede, O. J., Alaiyemola, F. F., & Okebukola, P. A. (1990). The effect of concept mapping on students' anxiety and achievement in biology. *Journal of Research in Science Teaching, 27* (10), 951-960.

Jolliffe, F. R. (2002). Sharing experiences in the training of researchers. *Statistics Education Research Journal, 1* (1), 14-15.

Kinchin, I. M., Hay, D. B., & Adams, A. (2000). How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development. *Educational Research, 42* (1), 43-57.

Klein, D. C. D., Chung, G. K. W. K., Osmundson, E., Herl, H. E., & O'Neil, J. F. (2002). *Examining the validity of knowledge mapping as a measure of elementary students' scientific understanding*. Center for the Study of Evaluation (CSE Tech. Rep. No. 557). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kottke, J. L. (2000). Mathematical proficiency, statistics knowledge, attitudes towards statistics, and measurement course performance. *College Student Journal, 34* (3), 334-347.

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., & Glaser, R. (2001). *Assessing student representations of inferential statistics problems* (CSE Tech. Rep. No. 553). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

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

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

McClure, J. R., Sonak, B., & Suen, H. K. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching, 36* (4), 475-492.

McKendree, J., Small, C., Stenning, K., & Conlon, T. (2002). The role of representation in teaching and learning critical thinking. *Educational Review, 54* (1), 57-67.

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press.

Novak, J. D. (1995). Concept mapping: A strategy for organizing knowledge. In S. M. Glynn and R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 229-245). Mahwah, NJ: Erlbaum.

O'Donnell, A. M., Dansereau, D. F., & Hall, R. H. (2002). Knowledge maps as scaffolds for cognitive processing. *Educational Psychology Review, 14* (1), 71-86.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88* (1), 144-161.

Quilici, J. L., & Mayer, R. E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Applied Cognitive Psychology, 16*, 325-342.

Rafferty, C. D., & Fleschner, L. K. (1993). Concept mapping: A viable alternative to objective and essay exams. *Reading Research and Instruction, 32* (3), 25-34.

Reynolds, R. E., Sinatra, G. M., & Jetton, T. L. (1996). Views of knowledge acquisition and repesentation: A continuum from experience centered to mind centered. *Educational Psychologist, 31* (2), 93-104.

Rice, D. C., Ryan, J. M., & Samson, S. M. (1998). Using concept maps to assess student learning in the science classroom: Must different methods compete? *Journal of Research in Science Teaching, 35* (10), 1103-1127.

Roberts, L. (1999). Using concept maps to measure statistical understanding. *International Journal of Mathematical Education in Science and Technology, 30* (5), 707-717.

Ruiz-Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (2001). Comparison of the reliability and validity of scores from two concept-mapping techniques. *Journal of Research in Science Teaching, 38* (2), 260-278.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33* (6), 569-600.

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

Schau, C., Mattern, N., Zeilik, M., Teague, K. W., & Weber, R. J. (2001). Select-and-fill-in concept map scores as a measure of students' connected understanding of science. *Educational and Psychological Measurement, 61* (1), 136-158.

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* (5), 484-494.

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

Silver, E. A. (1981). Recall of mathematical problem information: Solving related problems. *Journal for Research in Mathematics Education, 12* (1), 54-64.

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

Stice, C. F., & Alvarez, M. C. (1987). Hierarchical concept mapping in the early grades. *Childhood Education, 64* (2), 86-96.

Stoddart, T., Abrams, R., Gasper, E., & Canaday, D. (2000). Concept maps as assessment in science inquiry learning—a report on methodology. *International Journal of Science Education, 22* (12), 1221-1246.

Wells, F. B. (1999). The effect of the use of concept maps on community college students' conceptual understanding of biology course content. *Dissertation Abstracts International, 59* (7), 2433A. (University Microfilms No. AAM98-39954).

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 7* (3), 200-08, 210-214.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17*, 31-74.

APPENDIX A

STATISTICS PROBLEMS FOR SORTING TASK

Structural features	Surface features (cover story)		
Inferential test	Education	Weather	Politics
t-test	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. Problem 2	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 temperatures reported from the two locations varied. Temperature readings from both sites were recorded for one year. Problem 5	A political candidate wants to know whether voters' party affiliation (Democrat vs. Republican) varies by income level. The candidate's aide conducts a survey asking people to report their party affiliation and their total annual income Problem 7
Chi-square	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. Problem 1	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). Problem 10	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. Problem 12

Statistics Problems For Sorting Task (Cont.)

Inferential Test	Education	Weather	Politics
f-test	A professor wants to know when it is most useful to give students "organizers" that help structure math content. A group of 30 students are split into 3 groups of 10 each. The "organizer" is provided before learning new math content in Group 1 and after learning content in Group 2. Group 3 is not given an "organizer". All groups are then given a math test worth 25 points. Problem 4	A family plans to rent a cottage by the lake for their one-month summer vacation. Given that the weather has been highly variable in the last 5 years, the parents want to know which summer month tends to be consistently warm. They examine the temperatures during June, July, and August over the last five years to help them decide when to go on vacation. Problem 3	The Federal Communications Commission (FCC) has received complaints about many phone companies. One recurring complaint is the increasing high costs of service. Four companies were assessed over a 2-year period to determine which company's cost (in dollars) fluctuated less from month to month. This company was then recommended to customers. Problem 9
Correlation	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 T.V stifles creativity. Students' scores are recorded along with the reported number of hours of T.V they watch per week. Problem 11	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. Problem 8	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 states' health care ratings. Problem 6