Keywords: Knowledge structure; Selection skills; Statistical expertise; Statistical literacy; Statistical techniques.

Abstract

This study investigated the knowledge base necessary for choosing appropriate statistical techniques in applied research. In this study, we compared knowledge used by six experts and six novices in two types of statistical tasks. The tasks were: 1) comparing research scenarios from the perspective of choosing a statistical technique, and 2) direct comparison of statistical techniques. The framework was based on expert knowledge in inferential statistics using the repertory grid technique for data collection. A qualitative analysis of data showed that of the three types of expert knowledge, research design knowledge comprised the biggest portion, with theoretical and procedural knowledge comprising relatively smaller parts. Little difference was observed between experts and novices in extensiveness of knowledge use, although experts’ knowledge use was found to be more integrated than novices’. Finally, two implications were drawn regarding how to better teach selection skills in statistics education: (1) statistical techniques should be taught in relation to relevant research designs, and (2) conceptual connections between statistical techniques should be explicitly taught.

1. Importance of Selection Skills

Because coursework in statistics is a universal requirement for students pursuing doctoral dissertation research in the social sciences (among many other disciplines), having the appropriate selection skills (an individual’s ability to identify an appropriate statistical method or technique for a given research situation) in statistics is crucial. To achieve statistical sophistication, descriptive and inferential statistics are required investigative methodologies for understanding, evaluating and carrying out research at the
While students may “pass” some statistics courses by memorizing formulas and applying procedures to familiar, well-defined problems, they may not attain the knowledge required to model new problems using inferential statistics. Unfortunately, it is this very skill that is crucial when students do research in their own disciplines and engage in statistical data analysis (Committee on Undergraduate Programs in Mathematics, cited in Chervany, Collier, Fienberg, Johnson, and Neter 1977; Chervany, Benson and Iyer 1980). The need for students to reach this level of understanding has been documented as a necessary outcome of statistics education by both Gardner and Hudson (1999) and Quilici and Mayer (1996).

One factor contributing to this deficiency in understanding statistics is related to computational technology. Although user-friendly statistical software applications such as SPSS, MINITAB, and SAS can perform complex calculations in seconds, they can neither suggest nor select appropriate statistical techniques for given problems. This failure to model applied statistical situations correctly can easily produce a serious flaw --answers to the wrong questions (Tung 1989). This illustrates the point that different types of expertise are called into play during the course of statistical analysis. First is arithmetic expertise, the advanced computational ability of statistical software. Second is statistical expertise, which has to do with selecting appropriate statistical techniques and drawing sensible conclusions from the data (Hand 1984). This requires statistical literacy, a term coined by Watson and McGaw (1980) for the reasoning skills necessary to select appropriate statistical techniques in applied situations. Statistical literacy as a set of skills, includes recognition of statistics as an integral part of the empirical research process; comprehension of the (relatively few) organizing constructs in statistics; and awareness of the appropriate application and interpretation of various statistical techniques.

Within the field, there is a paucity of empirical research on statistical reasoning in general, and on statistical literacy in particular. This study seeks to help remedy that situation by comparing the knowledge of experts to that of novices, in an effort to establish the types of knowledge that statistical literacy is built upon. The findings of this study will be especially of interest to statistics educators seeking to help graduate students develop the selection skills necessary for inferential statistics.

2. Theoretical Framework and Research Questions

2.1 Theoretical Framework

In order to design sound instructional techniques that develop students’ selection skills in inferential statistics, it is important to investigate the factors that support these selection skills. Since inferential statistics experts already possess selection skills, it is reasonable to study expert knowledge in an effort to discover the nature of these supporting factors. In fact, considerable knowledge was found to be an essential basis for expert skill in several domains (e.g. Chi, Glaser and Rees 1982; Larkin, McDermott, Simon and Simon 1980).

Recent theories of human knowledge have pointed to the complex nature of expert knowledge in subject-matter domains (e.g., physics and medicine, which are comparable to inferential statistics in terms of sophistication). The term knowledge structure was coined to characterize the nature of an individual's knowledge in a domain (Chi and Koeske 1983). Knowledge structure denotes both the elements of what one knows in a domain, and how the elements are linked together. It accounts for the existence or absence of concepts within an individual's knowledge; the relationships between these concepts; and the degree to which the individual’s knowledge contains elements critical to performing valued tasks in a
domain (Chi and Koeske 1983). According to this theory, there are at least three characteristics that need to be considered while exploring knowledge structures -- content, organization and task-adaptedness.

This understanding of knowledge structures was derived from studies of experts and novices whose comparisons revealed patterns of differences within their knowledge. Knowledge structures of novices were found to be relatively poor in content (e.g., possessing fewer concepts and elements), and loose in organization (e.g., sparse or having fewer cross-references within its components) (Feltovich, Johnson, Moller and Swanson 1984).

Not surprisingly, there is evidence in the research literature pointing to differences in the knowledge structures of experts who specialize in performing different kinds of tasks within a given domain. When the goal of a given task is not specified, an expert who specializes in solving applied problems will attend to different aspects of a task than an expert who specializes in (re)producing theoretical generalizations of a discipline. That is, they will interpret the task in different ways, each one imposing a goal that is parallel to his or her area of specialization. For example, given a medical task, medical experts whose specialty is in scientific investigation of human pathophysiology, when compared to experts whose specialty is in clinical diagnosis, will attend to different aspects of the task, although their knowledge involves the same ontological entities. Based on this observation, it is postulated that their knowledge structures are adapted to different goals (Feltovich et al. 1984; Smith 1990; Smith 1992). We further postulate that this adaptation would show itself in the differing relative weights of the epistemological components of knowledge.

In this study, the nature of the knowledge base that can support selection skills in inferential statistics was investigated in relation to the three critical qualities of expert knowledge: content, organization and task-adaptedness. An individual’s knowledge was elicited and compared in tasks that required the use of such selection skills.

### 2.2 Research Questions

This study examined the following questions

1. How do experts with different professional specialities (statistical consultants and mathematical statisticians) compare in terms of the type of knowledge they use to perform statistical tasks that differ in goal-specificity? This question has three parts:
   
   1A. How does the *extensiveness* of the knowledge that the two kinds of experts use compare?
   
   1B. How does the *connectedness* of the knowledge that the two kinds of experts use compare?
   
   1C. How does the *task-adaptedness* of the knowledge that the two kinds of experts use compare?

2. How do experts and novices compare in the knowledge they use to perform statistical tasks that differ in goal-specificity? This question addresses two separate issues:

   2A. How does the *extensiveness* of the knowledge that experts and novices use compare?

   2B. How does the *connectedness* of the knowledge that experts and novices use compare?

The experts were chosen from two types of inferential statistics professionals: statistical consultants and mathematical statisticians. Statistical consultants were chosen because of their expertise in applied
problem solving; mathematical statisticians, because of their expertise in performing theoretical tasks.
The novices were graduate students who had completed courses in statistics, and thus still had relatively limited experience at using statistics.

Two types of task were used: one with a specific goal, and one in which a goal was not specified. The first task involved comparing and contrasting research scenarios in terms of applicable statistical techniques. This task had the specific goal of using statistics for applied problem solving. Five research scenarios were developed for this purpose. Scenarios were administered two at a time for comparison.

In the second task, participants were asked to compare statistical techniques in any way they could think of. The techniques chosen for interview were two-way ANOVA, chi-square test, test of Pearson’s product moment correlation coefficient $r$, and Sign test and t-test. The statistical techniques were presented in groups of three at a time for comparison. Subjects were interviewed using the repertory grid technique (Kelly 1955). This was the second type of task, using a context with an unspecified goal. This task allowed experts to impose a goal on the task in keeping with their area of expertise, i.e., applied problem solving for statistical consultants, and theoretical elaboration for mathematical statisticians.

Because the first task required attention to a specific goal of applied problem solving, knowledge used by both kinds of experts was expected to be similar in content and organization, and effectively adapted to solving applied problems. Knowledge used by statistical novices was expected to be relatively deficient in content (e.g., lacking components), and poor in organization (e.g., with missing or faulty connections). [Please note: we did not expect to find a difference between experts and novices in the task-adaptedness of knowledge use. Since the novices’ knowledge was not developed to the level that would reflect a primary orientation towards a particular goal in inferential statistics, we did not perform a comparison between experts and novices regarding this area.]

Research leads us to expect that tasks with an open-ended goal will lead individuals to impose an implicit goal in accordance with their experience and specialization (Smith 1990; Smith 1992). In this study it was expected that mathematical statisticians would attend to theoretical conditions and aspects of derivation of statistical techniques in the interview, whereas statistical consultants would attend more to pragmatic, research design features of situations in which these statistical techniques could be used. It was expected that statistical technique tasks would help reveal potential differences in the knowledge bases of the two kinds of experts, as a function of their imposition of a goal (applied problem solving or theoretical elaboration) due to their specialty. More specifically, the knowledge used by mathematical statisticians was expected to have more theoretical elements, whereas the knowledge used by statistical consultants was expected to have more research design knowledge. These differences were expected to be evident in the respective percentages of theoretical, and research design knowledge within an expert’s knowledge structure. They were also expected to give us insight into multiple expertise in the field of statistics. Such insight would be valuable because, since novices are educated to become “like experts,” it would be helpful to know what kind of experts we would like them to become in terms of the desired knowledge base.

3. Design of the Study

3.1 Subjects

The sample consisted of 12 subjects: 6 novices and 6 experts. Novices were chosen from among good standing doctoral students in mathematics education and in research methods. They had completed a series of three courses in statistics with a grade of B+ or better, but had relatively limited experience in using statistics.
The expert group was made up of two types of inferential statistics professionals: 3 statistical consultants chosen for their expertise in applied problem solving, and 3 mathematical statisticians chosen for their expertise in performing theoretical tasks.

We chose individuals with differing levels of statistical experience because experience is one factor that characterizes expertise in a domain (Reimann and Chi 1989).

3.2 Tasks Used in Data Collection

Two kinds of tasks for data collection were used: selecting appropriate statistical techniques for concrete research scenarios, and selecting names of appropriate statistical techniques in a context with a non-specified goal.

**Research Scenarios.** Five applied research scenarios were constructed. One of the five statistical techniques (listed below under “statistical techniques”) could reasonably be used to answer the research questions of these scenarios. The research scenarios reflected variations in the theoretical assumptions that needed to be satisfied for the desired statistical test (i.e., assumptions for parametric or nonparametric tests), and variations in the type of inference procedure used (testing hypotheses or building confidence intervals) as well as variations in five prominent features of research designs. These research design features were: (1) type of research questions (testing for difference or testing for association), (2) number of variables involved (bivariate or multivariate), (3) method of sampling (independent sampling or related sampling - matched/repeated measures), (4) level of measurement scale for dependent variables (nominal, ordinal, or interval/ratio), and (5) groups of independent variables (binomial or multinomial). The scenarios were reviewed and approved for face validity by two professors who taught graduate courses in inferential statistics. The research scenarios and the list of characteristics they were designed to reflect can be found in Appendix A.

Statistical Techniques. The five statistical tests chosen were: (1) Two-way ANOVA, (2) Test of Pearson's product-moment correlation coefficient ($r$), (3) Chi-square test, (4) T-test, and (5) Sign test. These five tests included both parametric tests (two-way ANOVA, t-test and test of Pearson's $r$ ) and non-parametric tests (Chi-square and Sign test). These statistical techniques were chosen on the basis of three criteria: (1) techniques had been covered in the courses novices had taken; (2) underlying theoretical aspects of the statistical tests varied; (3) research situations into which the statistical tests typically fit represented a variety of design features discussed above.

It is important to note here the delimitations posed by the tasks. For example, although ANOVA is the name of a general family of techniques, it was specified as “two-way ANOVA” in this study. The purpose was to elicit the individual’s knowledge in relation to particular research design specifications.

3.3 Procedures – Interview Using Repertory Grid Technique

Experts and novices were interviewed individually in three separate sessions, following the procedures of the Repertory Grid Technique (Kelly 1955). Constructs were elicited by using research scenarios in the first session, and by using the names of statistical tests in the second session. Constructs refer to statements of similarity or difference among research scenarios and among statistical techniques suggested by individuals. During the third session, subjects rated elements of both types (research scenarios and statistical techniques) based on the corresponding sets of constructs suggested in the two earlier sessions

**Construct Elicitation.** In the construct elicitation using research scenarios, the subject was first asked to read all five of the scenarios carefully. Then the research scenarios were given two at a time and the
subject was told, “assume that you are being asked to answer the research questions described in the scenarios. Answering the research questions involves some kind of statistical analysis. From the perspective of issues you consider for choosing an appropriate statistical technique, how are the two scenarios similar or different?” After responding, the subject was asked "are there any other ways that they are similar or different?”, until the subject indicated that he or she could not come up with any more aspects of similarity or difference for the pair under consideration. The process was repeated for the six dyads in which individual scenarios were represented with frequency in the pairs. Table 1 presents the dyads and the order they were given. The same set of six dyads was used for all subjects in the same order. The first session took 40 to 60 minutes.

| Dyad | First Scenario                                                                 | Second Scenario                                                                 |
|------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| Dyad 1 | Participation and achievement among 4th grade students. | Salaries of teachers with a master's degree. |
| Dyad 2 | Salaries of teachers with a master's degree. | Relationship between high school and college achievement. |
| Dyad 3 | Relationship between high school and college achievement. | Participation and achievement among 4th grade students. |
| Dyad 4 | Pairing in mathematics class. | Preparation time and success in examination. |
| Dyad 5 | Relationship between high school and college achievement. | Pairing in mathematics class. |
| Dyad 6 | Relationship between high school and college achievement. | Preparation time and success in examination. |

Although repertory grid technique typically involves giving three elements to be compared at a time, in this study, only two research scenarios were given at a time. This is because a pilot study had found that participants had difficulty considering three scenarios, as some were page-long texts. To alleviate the memory load, this slight change was adopted. Also, there were ten possible dyads of five scenarios. To keep the length of interview reasonable, only six dyads were used, but the dyads were formed in such a way that they included a similar number of scenarios.

In the second session, the names of statistical techniques were presented three at a time. Six triads of statistical techniques were chosen initially from among all ten possible triadic combinations of the statistical techniques. The purpose of using the six triads was again to keep the interview time short.
enough to be feasible. As in the case of pairs of research scenarios, the six triads were selected in such a way that individual statistical techniques were represented in the triads in similar numbers. The same six triads were used with all of the twelve subjects. Table 2 gives the triads and the order they were given.

Table 2. Statistical Technique Triads.

| Triad 1 | Triad 2 | Triad 3 | Triad 4 | Triad 5 | Triad 6 |
|---------|---------|---------|---------|---------|---------|
| Two-way ANOVA | Test of Pearson's $r$ | Two-way ANOVA | Test of Pearson's $r$ | Two-way ANOVA | Two-way ANOVA |
| X$^2$ Test | X$^2$ Test | Test of Pearson's $r$ | t-test | Test of Pearson's $r$ | X$^2$ Test |
| t-test | t-test | Sign Test | Sign Test | t-test | Sign Test |

The subjects were shown an index card with the names of three statistical techniques printed on it and were asked "How are these techniques similar or different?" Their responses were followed by the query, "are there any other ways that they are similar or different?" This question was repeated until the subject indicated that he or she could not come up with any more aspects of similarity or difference for the triad under consideration. As the subject talked about the similarities and differences, the interviewer wrote down the key phrases or words that were used to describe the distinctions. The construct elicitation sessions were tape-recorded to check the reliability of construct elicitation.

After each of the first and second sessions, the interviewer listened to the tape-recordings to check the completeness of the constructs in the notes taken during the interview. Two matrices were prepared for each subject, one for research scenarios and one for statistical techniques. The matrices had $N \times 5$ dimensions, with $N$ representing the number of constructs elicited from a particular subject for a given type of element (scenarios or statistical techniques). After eliminating those constructs that were repeated more than once, the constructs were transferred into rows of a matrix with five columns, in which the columns listed the elements (research scenarios in the first matrix, and statistical techniques in the second). A sample rating matrix for the statistical technique task with constructs and associated ratings completed by a novice subject is shown in Table 3 below.

Table 3. The constructs and the rating matrix elicited from Novice 3 by using the names of statistical tests.
|   | Two-way ANOVA | Test of Pearson's $r$ | $\chi^2$ Test | t-test | Sign Test |
|---|---------------|----------------------|--------------|--------|-----------|
| 1 | 5             | 1                    | 1            | 1      | 1         |
| 2 | 1             | 1                    | 5            | 1      | 1         |
| 3 | 1             | 5                    | 3            | 3      | 5         |
| 4 | 5             | 5                    | 3            | 1      | 1         |
| 5 | 5             | 1                    | 5            | 1      | 1         |
| 6 | 1             | 1                    | 5            | 1      | 1         |
| 7 | 1             | 1                    | 1            | 5      | 5         |
| 8 | 5             | 1                    | 1            | 1      | 1         |
| 9 | 1             | 1                    | 1            | 5      | 1         |
| 10| 5             | 5                    | 1            | 1      | 1         |
| 11| 1             | 1                    | 1            | 5      | 5         |
| 12| 1             | 1                    | 5            | 1      | 5         |
| 13| 1             | 5                    | 1            | 1      | 1         |
| 14| 1             | 1                    | 5            | 1      | 1         |
| 15| 1             | 1                    | 4            | 5      | 5         |
There was 89% agreement between the constructs recorded by the principal investigator and an independent reviewer on a subset of research scenario constructs from two subjects and a 94% agreement on statistical technique constructs. These reliability levels were considered satisfactory, and the disagreements were resolved for these cases. The remaining constructs were coded by the principal investigator.

In the third and final session, subjects read the lists of constructs to help them remember what they had suggested in the previous sessions, and checked to see if the wording reflected what they meant. Necessary changes were made if the subject so requested.

The subjects then indicated the extent to which they thought that a construct applied to each research scenario, or that it was true for the scenario, by completing a rating matrix. The subjects rated each construct using a Likert scale from 1 to 5.

The rating was explained to each subject as follows: "on a scale from 1 to 5, with 1 being the lowest and 5 being the highest, please rate the factors of similarity or difference you came up with, for how much they apply to each of the research scenarios. One indicates the factor applies very little or is not true for the scenario while 5 means it applies very much or is very true for the scenario. You can use the numbers in between (2, 3, and 4) accordingly." Rating of statistical techniques based on the corresponding constructs was done in a similar manner. Completing the rating matrices took 30 to 45 minutes.

The fact that research scenarios were used before statistical techniques in interviews reflected a choice between possible disadvantages of the "order effect." If the statistical techniques were used first, subjects might have noticed the obvious match between scenarios and statistical tests. This would have affected the nature and number of constructs elicited from the subjects. Giving the research scenarios first might have provided a similar "shadow" over the perception of the statistical technique task and lead subjects to see the techniques from particular matching research design perspectives. However, the adverse effects of this second factor were thought to be less probable and comparably less confounding, and thus the research scenarios were given first. The reader should read and interpret the findings with this delimitation in mind.

| 16. This test is taught in Stat I. | 1 | 5 | 1 | 5 | 1 |
|----------------------------------|---|---|---|---|---|
| 17. How different factors combine to affect another variable is what you can test for by using this technique. | 5 | 1 | 1 | 1 | 1 |
| 18. You can use this test with repeated-measures situations. | 5 | 1 | 1 | 1 | 1 |
| 19. When you have two different groups, this test may be used. | 5 | 1 | 5 | 3 | 5 |
| 20. This test can be used with randomized-block-type research designs. | 5 | 1 | 1 | 1 | 1 |
3.4 Data Analysis

Data. Two kinds of data were collected from experts and novices: (1) personal constructs describing the factors of perceived similarity or difference among the elements (research scenarios or statistical techniques), and (2) rating matrices reflecting the degree to which the constructs applied to scenarios and statistical techniques.

Two sets of constructs were obtained corresponding to the two types of elements (research scenarios and statistical techniques). The personal constructs were in the form of a sentence expressing an idea of similarity or difference. The personal constructs as they were obtained in the two tasks were considered to contain the components of knowledge used by individuals. Accordingly, analyses of the nature of constructs informed the content and nature of knowledge that an individual used to perform a statistical task.

Classification of Constructs. The constructs elicited from individuals were first classified in relation to the type of knowledge they represented based on a classification scheme developed Alacaci (1998). The constructs were classified into four categories: (1) research design aspects, (2) theoretical aspects, (3) procedural aspects, and (4) non-technical aspects.

Constructs in the research design category addressed a variety of characteristics of research situations. Research design constructs referred to features of research situations designed or manipulated according to the interests of the researcher before the statistical modeling of a research situation took place. The features falling into this category included purpose of analysis (comparison or correlation), numbers of independent and dependent variables, measurement scales of dependent and independent variables, number of categories of independent variable, types of relationship among samples, number of groups, population parameter, and other research design aspects (such as survey research, or experimental research).

Constructs related to theoretical aspects of statistical tests concerned the mathematical generalizations based on which the tests were created, and the ensuing theoretical conditions that had to be satisfied before the tests were used. These constructs pertained to required assumptions such as normality, equality of group variances, types of assumed relationships among variables (e.g., linear or nonlinear), conditions of sample size and considerations of power, and the type of underlying probability distributions.

Constructs classified into the category of procedural aspects of statistical knowledge were related to the procedures of the general paradigm of hypothesis testing. These constructs were related to a variety of issues including types of hypotheses to be tested (directional or non-directional), and the general computational aspects of the analysis process (such as computing degrees of freedom, values that test statistic take, meanings of computational procedures, etc.).

Constructs in the non-technical category referred to the personal opinions of subjects about the general qualities of statistical techniques or research scenarios. These did not fit into any of the categories listed above. Examples included perceived difficulty of the research scenarios, context of the scenarios, and popularity or usefulness of the statistical techniques.

These four categories of statistical knowledge served as a general framework for data analysis. Comparing subjects on the content of their knowledge base required, however, characterizing the underlying idea of each construct in order to describe them beyond the idiosyncratic wordings of individuals. This coding was also necessary to quantify knowledge falling into the four categories.

Thematic Coding of Constructs. Constructs were further coded for the themes they represented. A
Theme is a unitary idea or a concept that was addressed in the constructs. It was necessary to characterize constructs for the themes they represented for two reasons. First, subjects may express similar ideas in their constructs by using different terminology. In order to compare the constructs suggested by individuals, constructs were matched to a standard expression of the idea contained in them, i.e., the theme. Second, some constructs consisted of two ideas. In order to fully represent the content of a construct, these ideas were represented separately. In summary, themes provided a common framework upon which the constructs could be aggregated and compared across groups of individuals and across types of tasks.

After an inclusive list of themes addressed by all subjects in the two task environments was prepared, the themes addressed by a particular individual could be determined by checking the actual themes observed in his or her set of constructs. Since the list of the themes was all-inclusive, the themes addressed by an individual were necessarily a subset of the complete list. The complete list of the themes addressed in the constructs of individuals can be found in Appendix B.

For consistency in thematic coding, the lists of themes coded by an independent reviewer and the principal investigator on constructs elicited from a novice and an expert subject were compared. An 86% agreement was observed for the constructs from the novice subject and a 97% agreement was observed for the expert subject. These levels of agreement were considered satisfactory, and the remaining constructs were coded by the principal investigator only.

Correctness of Themes. A given theme addressed in an individual's construct was not always correctly related to all of the elements of a task (research scenarios or statistical techniques). The correct application of the themes was evaluated from the matrices of ratings. For example, as seen in Figure 3, novice 3 rated t-test as a technique that could not be used for repeated-measure situations. This subject had a limited perception of the situations in which t-test could be used with regard to relationships among samples.

The number of themes addressed and the number of correctly-applied themes in each category of knowledge were scored separately for the two tasks for a given individual.

Measures of Knowledge. In this study, we wanted to compare knowledge used by individuals in the tasks. Comparisons were performed on three measures: (1) extensiveness, (2) connectedness, and (3) task-adaptedness of the knowledge used to perform the tasks.

Extensiveness of knowledge used by individuals was defined as the combined number of themes an individual addressed in the three categories: research design, theory and procedures. Since themes denoted ideas contained in the elicited constructs, a higher number of themes addressed by an individual represented a more comprehensive set of ideas used while performing the tasks. Hence, number of themes was a reasonable measure of the extensiveness of an individual’s knowledge. Themes in the non-technical category were not considered as part of the extensiveness of an individual’s knowledge. This type of knowledge was personal in nature, was not principled, and was not equally applicable to both task environments.

Connectedness of knowledge was derived from an evaluation of an individual’s rating matrix. A given theme of knowledge addressed by an individual in a construct was evaluated for its correct application to the scenarios and the statistical techniques.

A less-developed knowledge structure could vary from a well-developed knowledge structure in two different ways. First, the number of themes that the knowledge structure comprised may have been fewer in number, and hence the knowledge structure may have been less “extensive.” Second, for the existing
themes of knowledge, the less-developed knowledge structure may have consisted of even fewer themes that were correctly associated (or correctly not associated) with the statistical techniques. For example, t-test could be used for situations in which the categories of independent variable are binomial (two categories). If the situation were multinomial, then an ANOVA would be used. Thus, if an individual rated t-test as a multinomial test, this was considered an “incorrect association.”

It is important to note that not every theme applied equally to all the scenarios or statistical techniques. For example, the research design criteria of the number of groups of independent variables did not apply to situations where Test of Pearson’s moment correlation coefficient could be used. When an expert or a novice rated the Test of Pearson’s $r$ or the corresponding scenario for this theme as “1”, it was considered “correct association” for the connectedness measure of knowledge.

The combined number of themes in research design, theoretical and procedural categories that were correctly associated (or correctly not associated) with all five elements for a given type of task was taken as a measure of the connectedness of an individual's knowledge structure in this study.

For task-adaptedness, it was assumed that an individual’s knowledge may be adapted to one of the two common goals in statistics: applied problem solving, or theoretical (re)production of statistical knowledge. It was conjectured that elements of knowledge related to research design supported the applied problem solving tasks, whereas theoretical elements supported the theoretical tasks (Alacaci 1998). Hence, the percentage of the number of research design themes and the percentage of theoretical themes within the overall number of themes addressed by an individual were used to define how much one’s knowledge was adapted to an applied problem solving goal and a theoretical goal, respectively.

**Methods of Analyses.** For the first research question, the two kinds of experts (statistical consultants and mathematical statisticians) were compared for the three measures of knowledge: extensiveness, connectedness, and task-adaptedness. For the second research question, the knowledge of experts and novices was compared only on the extensiveness and connectedness.

There were no compelling theoretical reasons to expect differences in task-adaptedness between experts and novices. Task-adaptedness of one’s knowledge is attained through extensive experience of using knowledge for a particular goal (Smith 1990). As novices lacked the level of experience to specialize in statistics, experts and novices were not compared on this measure.

Because the small sample size did not allow performing statistical tests on the parameters of interest, only descriptive and qualitative comparisons are reported.

**4. Results**

Comparisons for the two tasks (research scenarios and statistical techniques) are reported separately.

**4.1 Comparison of Statistical Consultants and Mathematical Statisticians**

Research scenario tasks exemplify applied problem solving goals. Although statistical consultants and mathematical statisticians had different specializations, both groups were expected to have well-developed and dynamic knowledge bases to perform the research scenario tasks. Accordingly, it was conjectured that no difference would be observed in the extensiveness, connectedness or task-adaptedness of the knowledge used by the two groups of experts.

In statistical technique tasks, knowledge used by the two kinds of experts was again expected to be
similar in the extensiveness and connectedness measures of knowledge, because these measures were related to level of experience as the two groups of experts were similar in the level of their experience in inferential statistics. The knowledge experts used was expected, however, to differ in task-adaptedness, because it involved comparing statistical techniques in an abstract context. When experts were asked to perform tasks with an unspecified goal, they were expected to impose a goal that was parallel to their specialization (Smith 1990; Smith 1992). Statistical consultants were expected to impose an applied problem solving goal to the statistical technique tasks, while mathematical statisticians were expected to impose a theoretical goal parallel to the way they used statistics most often.

More specifically, themes of knowledge related to research design aspects of statistical techniques were expected to occupy a bigger percentage of the knowledge used by statistical consultants compared to that of mathematical statisticians. On the other hand, a difference in favor of mathematical statisticians was expected for theoretical themes, since theoretical elements of knowledge facilitated theoretical tasks most.

**Extensiveness and Connectedness of Knowledge:** Figure 1 presents information about extensiveness of knowledge used by the two groups of experts in research scenario tasks. Figure 2 presents similar information for connectedness.

![Figure 1. Extensiveness of the knowledge used by statistical consultants and mathematical statisticians in research scenario tasks.](image)

In Figures 1-6, the bars represent individual measures of knowledge and group averages. “ExpSC1” stands for the first statistical consultant expert, “ExpSC2” for the second statistical consultant. “ExpMS1” represents the first mathematical statistician expert, and the rest of the individual experts are identified accordingly. Also, “AverageExp.Con.s” identifies the bar indicating the average value for the expert consultant group, while “AverageExpM.Stat.s” shows the similar measure for the mathematical
statistician expert group.

Figure 1 does not suggest evidence of a difference between the two groups, once the trend of differences within groups is considered. This leads us to propose that there does not seem to be a difference between the extensiveness of the knowledge used by the two groups of experts in the research scenario task.

![Figure 2](image)

Figure 2. Connectedness of the knowledge used by statistical consultants and mathematical statisticians in research scenario tasks.

There is a small difference between the mean number of themes correctly used by statistical consultants and those used by mathematical statisticians, as Figure 2 shows. Again, the figure does not reveal a trend suggesting a difference in connectedness of their knowledge use when considered in the context of the distribution of the measure within groups.

Figures 3 and 4 show the extensiveness and the connectedness measures of knowledge in statistical technique tasks respectively. It is clear that mean differences between the two groups are relatively small and the data do not reveal a difference between the two groups in the extensiveness and connectedness measures of knowledge use by the two groups of experts, taking into account within group variability.
Figure 3. Extensiveness of knowledge used by the two groups of experts in statistical technique tasks.

Figure 4. Connectedness of knowledge used by the two groups of experts in statistical technique tasks.

Task-adaptedness of the Knowledge: Figure 5 displays the percentages of the research design,
theoretical and procedural themes addressed within the knowledge used by experts in research scenario task. Figure 6 shows similar information for the statistical technique task.

Figure 5. Task-adaptedness of knowledge used by the two groups of experts in research scenario tasks.

Figure 6. Task-adaptiveness of knowledge used by the two groups of experts in statistical technique tasks.
Figures 5 and 6 show that the biggest portion of the knowledge used by experts in both types of tasks is comprised of research design knowledge (with the exception of one expert, ExpertMS1, in the statistical technique task). It is also observed that when Figures 5 and 6 are considered together, the percentages of theoretical themes seem to occupy a bigger portion within the knowledge used by mathematical statisticians in statistical technique tasks than the knowledge used by the same group in research scenario tasks. For statistical consultants, on the other hand, the percentage of research design knowledge is consistently high in both types of tasks.

A closer examination reveals that relative differences between the two expert groups in the mean percentages of research design and theoretical themes are in predictable directions: The mean percentages of research design themes within statistical consultants’ knowledge use are 80% and 77% in the research scenario and statistical technique tasks. The same measures for mathematical statisticians are 67% and 49%, which are both lower than that of statistical consultants. The trends in the data within groups suggest that consultants’ knowledge use is better adapted to applied problem solving than is the knowledge of mathematical statisticians, and this trend is more pronounced in the statistical technique task.

Similarly, the mean percentages of theoretical themes within mathematical statisticians’ knowledge use are 26 % and 40 % in the research scenario and statistical technique tasks respectively. The measures for consultants are 16 % and 11 %. In the mean percentage of theoretical themes, mathematical statisticians have a 10 % lead in research scenario tasks and a 29 % lead in the statistical technique tasks. In other words, mathematical statisticians’ knowledge use appears to be more adapted to theoretical tasks than consultants’ knowledge use, and this is more evident in the statistical technique tasks.

To summarize, the findings of this study tend to suggest that there is no difference in the extensiveness and connectedness measures of knowledge use between the two expert groups in any of the two types of tasks. However, statistical consultants’ knowledge use seems more adapted to applied tasks than mathematical statisticians’ knowledge use. Similarly, mathematical statisticians’ knowledge use seems to be more adapted to theoretical tasks. These generalizations are more readily observable in the statistical technique task in which a goal is not specified.

4.2 Comparison of Novices and Experts

Differences were expected between experts and novices in the extensiveness, and connectedness of the knowledge use in both tasks environments. Compared to experts, it was anticipated that the knowledge used by novices would be sparse in content, and less inter-connected.

In Figure 7, experts are indicated by Expert SC1 for the first statistical consultant expert, Expert MS2 for the second mathematical statistician, and Novice1 for the first novice, etc. The chart reveals a difference between the mean number of themes addressed by experts and novices is 6.1 between the mean number of themes addressed by experts and novices, favoring experts. Although the difference is in the expected direction when variability of data within groups is considered, it does not seem to be large enough to suggest a difference in the extensiveness of knowledge used by experts and novices.
Figure 7. Extensiveness of knowledge used by experts and novices in research scenario tasks.

There is a difference of 12.2 in the mean number of themes correctly associated with research scenario tasks by experts and novices, as Figure 8 shows. When variability within groups are considered, a difference between the two groups in the connectedness of knowledge use can be suggested in this type of task.

Figure 8. Connectedness of knowledge used by experts and novices in statistical technique tasks.
Figure 9 shows that experts addressed 4.7 more themes than the novices, on average, in statistical technique tasks. When variability of the measure within groups are taken into consideration, the difference does not seem to be large enough to suggest that experts’ knowledge use is more extensive than that of novices’ in this type of task.

Figure 9. Extensiveness of knowledge used by experts and novices in statistical technique tasks.

Figure 10 displays a difference between experts and novices of 12.6 themes, regarding themes correctly associated with statistical techniques between experts and novices. This difference suggests that experts’ knowledge use is more connected than novices’ knowledge use in statistical technique tasks when within group variability is considered.
To summarize, when experts were compared to novices, there was insufficient evidence to suggest a difference in the extensiveness of the knowledge used, not supporting the theoretical expectations at the start of the study. However, the findings did suggest that knowledge used by experts is more interconnected in both the research scenario and statistical technique tasks, which supports the preliminary expectations.

5. Further Analysis and Discussion on the Nature of Knowledge That Can Support Selection Skills in Inferential Statistics.

Although some differences were observed between the two groups of experts in the level of adaptation to applied problem solving and theoretical goals, it is important to note that the knowledge used by both groups was still primarily adapted to the applied goal as elicited in the tasks of this study. The biggest portion of the knowledge used by the two expert groups in both the research scenario and statistical technique tasks can still be accounted for by research design themes. This has implications for understanding the nature of well-developed knowledge bases that can support selection skills in inferential statistics.

The nature of well-developed knowledge structures that can support selection skills can also be informed by looking into the themes that are addressed by a majority of individuals in expert and novice groups (n≥4). Appendix B presents the themes addressed by all the participants, with the themes addressed by a majority of experts given in bold, and the themes addressed by a majority of novices given in italics. Table 3 provides the number of themes addressed by the respective majorities in the expert and novice groups.
Of the 21 themes addressed by a majority of experts, 17 belonged to the research design category, three were theoretical, and one was procedural. This finding further corroborates the observation about the importance of research design knowledge in well-developed knowledge structures needed for statistical modeling of research situations. It is important to note that only 11 research design themes were addressed by a majority of novice subjects. This, when considered with the data on the experts, shows that knowledge used by experts is richer in research design elements compared to novices’ knowledge use.

Further, the finding regarding novices’ knowledge being less-interconnected highlights the need to establish conceptual connections between statistical techniques. To develop expert-like knowledge bases, it is probably not sufficient for novices to learn about generic concepts of inferential statistics. It is necessary to make explicit the connections across the techniques regarding how these concepts apply (or do not apply) to specific techniques. The seeming lack of difference in the extensiveness of knowledge implies that novices do succeed to learn a great deal of generic concepts in inferential statistics. However, it appears that knowledge of how these concepts apply to specific techniques needs to be developed.

The findings of this study provide important insights regarding experts’ and novices’ knowledge, and about the factors that can support selection skills in inferential statistics. The findings should however, be interpreted in light of the limitations of the study. First, because of the small sample size, this study can only provide preliminary evidence. We recommend that the reader considers replicating the study, using a larger sample to allow traditional statistical testing for group comparisons. This will afford definitive evidence to support or refute the conjectures brought up and discussed in the study.

6. Implications For Statistics Education

The analysis in this study of the differences and similarities between expert knowledge and novice knowledge can help statistics educators design instructional experiences that will be effective in developing the ability to select appropriate statistical techniques.

Two implications can be drawn for inferential statistics education based on the results of this study. First, statistical techniques should be taught in relation to the features of research design with which they can be used. Second, statistics education should stress the potential conceptual connections (that is, the similarities and differences) among the techniques with regard not only to the associated research design features, but also to the theoretical and procedural aspects.

6.1 Teaching Statistical Techniques in Relation to Research Design Features

Based on the finding that research design knowledge constitutes the backbone of a well-developed knowledge base about statistical techniques, it can be suggested that instruction in inferential statistics should explicitly aim to teach the research design aspects in order to better foster an expert-like knowledge base. Teaching research design features of statistical techniques can be done in a variety of

Table 3. Number of research design, theoretical and procedural themes addressed by a majority in expert and novice groups.

| Research Design Theory Procedures | Experts | Novices |
|----------------------------------|---------|---------|
|                                 | 17      | 11      |
| Theory                          | 3       | 4       |
| Procedures                      | 1       | 1       |

Of the 21 themes addressed by a majority of experts, 17 belonged to the research design category, three were theoretical, and one was procedural. This finding further corroborates the observation about the importance of research design knowledge in well-developed knowledge structures needed for statistical modeling of research situations. It is important to note that only 11 research design themes were addressed by a majority of novice subjects. This, when considered with the data on the experts, shows that knowledge used by experts is richer in research design elements compared to novices’ knowledge use.
One way to teach research design aspects is to make general associations with the techniques and the aspects of research design. When a specific statistical technique is being covered, the instructor can point out the important components of the research design aspects in relation to the technique, and organize a list of the characteristics of the technique in the form of an I.D. (identification) of the statistical technique. This I.D. can include the specific aspects of research design that are always true for the technique (e.g., t-test can be used with data in interval or ratio scales) as well as the aspects that can vary (e.g., t-test can be used with matched-pairs, repeated-measures and independent samples). The I.D. of the test can also include theoretical conditions (e.g., for t-test, normal distribution of data and equality of variances) as well as procedural aspects (e.g., t-test can test one-directional as well as non-directional hypotheses). The I.D. of the techniques can potentially include elements corresponding to themes, such as those (see Appendix B) addressed by a majority of the experts in this study. These themes include research design aspects pertaining to the purpose of the test, the number and kinds of variables, the measurement scale of dependent variables, the relations among samples, the number of groups, and population parameters. In addition, the instructor may modify the elements of the I.D. depending on the kinds of techniques covered in class. Since the structural features (i.e., the research design aspects) of applied problem situations in inferential statistics are relatively few in number, it is feasible to teach these features.

Another strategy for instructors is to “think aloud” while leading students through choosing appropriate statistical techniques for applied research situations. This will make the decision making process of instructors visible (and audible) and help students see that it is not magic but built from an existing knowledge grid that they can learn to mimic as they make their own decisions in similar situations. The benefits of such processes have been documented by statistics educators to improve students' selection skills (Ware and Chastain 1991; Kenney 1998).

The third strategy for teaching research design aspects of statistical techniques is by using actual examples of research (such as published research in journals). In fact, this strategy has long been advocated by statistics educators (e.g., ASA 1980; 1982; Easton, Roberts, and Tiago 1988; Hogg 1991; Snee 1993).

Finally, using research scenarios that are carefully constructed by instructors to fit the specific situations of statistical techniques, can also improve the learning of research design aspects of the techniques. The scenarios can be used to introduce the technique initially, or to elaborate on the varied research situations (in addition to the most typical situations) with which the techniques can be used. These scenarios can be brief, and they can illustrate more directly the point the instructor wishes to make with respect to research design aspects, thereby minimizing contextual features which could potentially confuse or distract the students.

6.2 Teaching Conceptual Connections Among Statistical Techniques

Building conceptual connections among statistical techniques is of crucial importance in helping novices develop selection skills. Choosing a statistical technique often requires an initial consideration of a candidate technique. When one of the conditions of the technique is not met by the research situation under consideration, the individual may have to switch from the original technique to another one. This ability to switch to another technique can be aided by knowing how techniques are connected conceptually. The connections between techniques are provided by the themes of knowledge, as these themes can serve as the tools with which to describe the techniques as being similar or different. The conceptual connections among the techniques can be built by teaching the techniques while making explicit the connections between them.
As statistical techniques are taught, students can be asked to describe the similarities and differences between the new technique and the techniques that have already been covered. This can take the form of an assignment, be part of class instruction, or even be a class project for the assessment of student learning. For example, students could be asked to create a decision tree or flowchart to represent the relationship between techniques schematically. The flowchart could address the critical research design aspects as well as the required theoretical assumptions and the associated procedural aspects. The wider the scope of techniques that the flowchart covered, the more beneficial the exercise would be, because more connections would be addressed. Another version of this technique would be to use small-scale flowcharts or "micro-trees". These could be developed initially for specific families of techniques (such as t-tests with equal, and non-equal sample sizes; or t-tests with different types of relations among samples - independent, repeated measures, matched-pairs, etc.), and later integrated into a larger composite chart.

An interesting use of this strategy is illustrated by Chervany, et al. (1980) As part of the course requirements, they asked graduate students taking an introductory inferential statistics class to create a decision tree, displaying how to select the techniques covered in the class. Chervany and his colleagues found positive correlation between the score on the decision tree assignment and the score on the final exam questions, where students were required to select techniques for applied research situations. They also commented that students' work on decision trees provided information for instructors about the points where the students had difficulty (Chervany, et al. 1980).

Different versions of this strategy can be used, in accordance with the purposes of instructors, to help the novices develop connections among techniques. For example, after a chapter is covered and the corresponding decision trees are created by the students, all or a subset of students' decision trees can be discussed in the class, and a "correct"; decision tree can be provided to the students at the end.

Acknowledgements

This research is based on a dissertation study (Alacaci 1998) by the author at the University of Pittsburgh, School of Education, under the guidance of Dr. Ellen Ansell. The author gratefully acknowledges her mentorship and support. The help of committee members Drs. Edward Silver, Martin Cohen and Louis Pingel is also appreciated. This research was supported by a grant from the Faculty and Student Research Award of the School of Education at the University of Pittsburgh. The author would also like to thank Francis Di Vesta, Tom Short and the two anonymous reviewers whose comments improved the quality of this paper significantly.

References

ASA (American Statistical Association) (1980), “Preparing statisticians for careers in industry,” The American Statistician, 34, 65-80.

ASA (American Statistical Association) (1982), “Preparing statisticians for careers in federal government,” The American Statistician, 36, 69-89.
Alacaci, C. (1998), "Investigating Knowledge Structures in Inferential Statistics," unpublished Ph.D. Dissertation, University of Pittsburgh, College of Education.

Chervany, N.L., Collier, R.O., Fienberg, S.E., Johnson, P.E., and Neter, J. (1977), “A framework for the development of measurement instruments for evaluating the introductory statistics course,” *The American Statistician, 31*, 17-23.

Chervany, N.L., Benson, G. and Iyer, R.K. (1980), “The planning stage in statistical reasoning,” *The American Statistician, 34*, 222-226.

Chi, M.T.H., Glaser, R., and Rees, E. (1982), in “Expertise in problem solving,” *Advances In The Psychology Of Human Intelligence* (p.7-75), ed. R. Sternberg, Hillsdale, NJ: Erlbaum.

Chi, M.T.H., and Koeske, R.D. (1983), “Network representation of a child's dinosaur knowledge,” *Developmental Psychology, 19*, 29-39.

Curtis, D.A., and Harwell, M. (1996), “Training graduate students in educational statistics,” Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Easton, G., Roberts, H.V., and Tiao, G.C. (1988), “Making statistics more effective in schools of business,” *Journal of Business and Economics Statistics, 6*, 93-98.

Feltovich, P.J., Johnson, P.E., Moller, J.H., and Swanson, D.B. (1984), “LCS: The role and development of medical knowledge in diagnostic expertise,” in *Readings In Medical Artificial Intelligence: The First Decade* (pp. 275-319), eds. W.J. Clancey, and E.H. Shortliffe, Reading, MA: Addison-Wesley.

Finn, J.L., and Cox, D. (1992), “Participation and withdrawal among fourth grade pupils,” *American Educational Research Journal, 29*(1), 141-162.

Gardner, P.L., and Hudson, I. (1999), “University students’ ability to apply statistical procedures,” *Journal of Statistics Education* [Online] 7(1). [ww2.amstat.org/publications/jse/secure/v7n1/gardner.cfm](http://ww2.amstat.org/publications/jse/secure/v7n1/gardner.cfm)

Hand, D.J. (1984), “Statistical expert systems: necessary attributes,” *Journal of Applied Statistics, 12*(1), 87-112.

Hogg, R.V. (1991), “Statistics education: improvements are badly needed," *The American Statistician, 45*, 342-343.

Kelly, G. (1955), *The Psychology of Personal Constructs*, New York, NY: Norton.

Kenney, P.A. (1998), “Teaching introductory statistical methods for understanding: techniques, assessment and affect,” Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Larkin, J.H., McDermott, J., Simon, D.P., and Simon, H.A. (1980), “Expert and novice performance in solving physics problems,” *Science, 208*, 1335-42.

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

Reimann, P., and Chi, M.T.H. (1989), “Human expertise,” in *Human and Machine Problem Solving*
Smith, M.U. (1990), “Knowledge structures and the nature of expertise in classical genetics,” Cognition and Instruction, 7(4), 287-302.

Smith, M.U. (1992), “Expertise and the organization of knowledge: unexpected differences among genetic counselors, faculty and students on problem categorization tasks,” Journal of Research in Science Teaching, 29(2), 179-205.

Snee, R.D. (1993), “What is missing in statistical education?” The American Statistician, 47, 149-154.

Tung, S.T.Y. (1989), An expert system for ANOVA based on a study of the statistical consulting process, unpublished Ph.D. dissertation, University of Delaware.

Ware, M.E., and Chastain, J.D. (1991), “Developing selection skills in inferential statistics,” Teaching of Psychology, 14(4), 219-222.

Watson, G., and McGaw, D. (1980), Statistical Inquiry, New York, NY: John Wiley and Sons.

---

Appendix A

**RESEARCH SCENARIOS**

*Please Note:* The specific names of statistical techniques appropriate for the scenarios and the research design features of the scenarios listed below were not part of the tasks given to subjects.

---

Scenario for two-way ANOVA

| participation group | Grade 1 | Grade 2 | Grade 3 |
|---------------------|---------|---------|---------|
| mean (s.d.)         | mean (s.d.) | mean (s.d.) |
Are there significant differences among the mathematics test scores of nonparticipating, passively participating, and actively participating 4th grade students? Is there a significant interaction between achievement of students at the three grade levels and the participation status at the 4th grade?

† Since the mathematics skills raw score distributions from the standardized test were highly skewed to the left, these were transformed to $x^* = 10 - \ln(max + 1 - x)$ where $x$ is the original score, $max$ the maximum total score on the test, and $x^*$ the transformed score used in the analysis. The distributions of transformed scores ($x^*$) were almost symmetrical and resembled a normal distribution. The three groups' covariance matrices were approximately equal, and were averaged. The pooled covariance matrix showed compound symmetry.

I.D. OF THE TASK

type of research question: test of difference
number of variables: three (multivariate)
method of sampling: repeated measures
measurement of dependent variable: interval
groups of independent variable: multinomial
theoretical assumptions: (1) normality, (2) homoscedasticity assumption, (3) independence of group observations.
type of inference procedure: two-tailed hypothesis testing

* adapted from Finn and Cox 1992.

Scenario for test of Pearson's r

THE RELATIONSHIP BETWEEN HIGH SCHOOL AND COLLEGE ACHIEVEMENT

For the last two years, students were being admitted to a large university on the basis of their high school grades. The Dean of Admissions wondered how strong the relationship was between high school grades and college grades. He randomly sampled 60 seniors from the university and obtained their high school and college GPAs (both out of 4.00. The following data were obtained after the GPA scores were rounded to the nearest tenth digit.

| Subject | 1  | 2  | 3  | ...... | 59 | 60 |
|---------|----|----|----|-------|----|----|
| non-participants | 8.50† | (.83) | 9.15 | (.89) | 9.72 | (.95) |
| passive participants | 8.76 | (.79) | 9.50 | (.78) | 10.07 | (.82) |
| active participants | 9.11 | (.77) | 9.95 | (.73) | 10.51 | (.67) |
The standard deviations of high school and college GPAs are 1.2 and 0.9, respectively for the sample. When the frequencies of high school and college GPAs were depicted separately on graphs, both had an approximate bell shape. Based on the data, is there enough evidence to suggest that students with higher high school GPAs tend also to have higher college GPAs (use alpha=0.01)?

I.D. OF THE TASK

**Type of research question:** test of association  
**Number of variables:** two (bivariate)  
**Method of sampling:** repeated measures  
**Measurement of dependent variable:** interval  
**Groups of independent variable:** N.A  
**Theoretical assumptions:** 1) bivariate normality, 2) homoscedasticity assumption.  
**Type of inference procedure:** one-tailed hypothesis testing

Scenario for 3 x 2 Chi-square test

### PREPARATION TIME AND SUCCESS IN EXAMINATION

The following table shows the number of students who passed or failed an examination, and the number of hours that these students dedicated to the study of topics covered in the exam. After the final exam, students were asked how much time they devoted for the preparation of the exam, and were classified into three groups (less than 5 hours, between 5 and 10 hours, and more than 10 hours). Following are the results.

|                | Pass | Fail | Total |
|----------------|------|------|-------|
| Less than 5 hours | 11   | 15   | 26    |
| 5 to 10 hours    | 25   | 7    | 32    |
| More than 10 hours | 35   | 7    | 42    |
| Total            | 71   | 29   | 100   |

Based on the information above, can you conclude that there is a statistically significant difference in success between the groups of students with different numbers of hours of preparation?
I.D. OF THE TASK

type of research question: test of difference
type of inference procedure: two-tailed hypothesis testing
number of variables: two (bivariate)
method of sampling: independent samples
measurement of dependent variable: nominal
groups of independent variable: multinomial
theoretical assumptions: (1) within group independence.

Scenario for t-test

SALARIES OF TEACHERS WITH A MASTER'S DEGREE

The teachers' professional development committee of the Pennsylvania Union of Teachers undertakes a study to find out whether having a master's degree is financially rewarding for elementary school teachers. Elementary school teachers with terminal bachelor's and terminal master's degrees (teachers who do not have degrees higher than bachelor's and master's degrees, respectively) from suburban school districts were chosen for the study. Only teachers with 10 to 15 years of experience were included. One teacher was selected per school. Thirty five teachers with a master's degree, and 35 teachers with a bachelor's degree were randomly selected. Information about their annual salary was collected by reviewing school records, after securing permission for it. Following are the results.

|                           | Average annual salary | standard deviation |
|---------------------------|-----------------------|--------------------|
| Teachers with a bachelor's degree | $32,100               | $5,100             |
| Teachers with a master's degree | $36,475               | $4,400             |

Based on the findings of the study above, is there enough evidence to conclude that, on average, elementary school teachers with a master's degree earn more money than those with a bachelor's degree (use alpha=0.05)?

I.D. OF THE TASK

type of research question: test of difference
number of variables: two (bivariate)
method of sampling: independent
measurement of dependent variable: interval
groups of independent variable: binomial
theoretical assumptions: (1) normality, (2) equality of group variances, (3) independence of group observations.
type of inference procedure: one-tailed hypothesis testing

Scenario for Sign Test

PAIRING IN MATH CLASS

To test the effects of mixed-ability grouping on achievement, Mrs. White formed 14 pairs of one less-able and one more-able student in her 4th grade mathematics class in September. One student from below and one from above the median based on the last year's grade reports in math were randomly assigned to each other to make a pair. The teacher wanted to find out which group of students (less-abled students and more-abled students) benefit more from pairing. The pairs worked together while solving arithmetic word problems. To make sure that more-able students did not do all the talking in the pairs, they were instructed to take turns explaining the problem to each other. Inter-pairs communication was not allowed during the problem solving sessions. The experiment lasted for four and a half months, students working in pairs one class-hour a week. Students were administered two parallel forms of a standardized problem solving test before and after the experiment. The first score was subtracted from the second to find the Improvement Score for each student. Then the Improvement Scores of the members of pairs were compared. If the less-able student made more progress than the more-able student, the pair was assigned a "+". If the reverse was true, than the pair was assigned a "-". If they made equal improvement, the pair was assigned a "0". Find the 95 percent confidence interval to test whether less-abled students benefited from pairing in the math class.

I.D. OF THE TASK

type of research question: test of difference
number of variables: two (bivariate)
method of sampling: related (matched-pairs)
measurement of dependent variable: ordinal
groups of independent variable: binomial
theoretical assumptions: (1) variable based on which the "signs" were assigned has a continuous distribution.
type of inference procedure: building a confidence interval

Appendix B

Themes Addressed in the Elicited Constructs of Subjects
Table 4. Research design themes observed in the elicited constructs and the number of subjects who addressed the themes.

| Themes                                                                 | Experts (n=6) | Novices (n=6) |
|-----------------------------------------------------------------------|---------------|---------------|
|                                                                       | research      | statistical    | research | statistical | scenarios | scenarios | scenarios | scenarios |
| PURPOSE                                                               |               |               |           |           |           |           |           |           |           |
| 1. test of difference*                                                 | 6             | 5             | 6         | 4         |           |           |           |           |
| 2. test of association                                                 | 5             | 6             | 5         | 6         |           |           |           |           |
| 3. goodness of fit test                                                | 0             | 2             | 0         | 0         |           |           |           |           |
| 4. testing homogeneity                                                 | 2             | 2             | 0         | 0         |           |           |           |           |
| NUMBER OF INDEPENDENT, DEPENDENT AND ALL VARIABLES                    |               |               |           |           |           |           |           |           |
| 5. all, one variable                                                  | 0             | 1             | 1         | 0         |           |           |           |           |
| 6. all, two variables                                                 | 5             | 5             | 2         | 4         |           |           |           |           |
| 7. all, three variables                                               | 1             | 2             | 1         | 1         |           |           |           |           |
| 8. independent, one variable                                          | 2             | 2             | 0         | 2         |           |           |           |           |
| 9. independent, two variables                                         | 4             | 4             | 3         | 4         |           |           |           |           |
| 10. main effects                                                       | 3             | 1             | 0         | 2         |           |           |           |           |
| 11. interactions                                                      | 3             | 4             | 5         | 3         |           |           |           |           |
| 12. test of association                                               | 1             | 3             | 0         | 3         |           |           |           |           |
| MEASUREMENT SCALE OF INDEPENDENT VARIABLE                             |               |               |           |           |           |           |           |           |
| 13. nominal                                                           | 1             | 3             | 1         | 1         |           |           |           |           |
| 14. ordinal                                                           | 2             | 1             | 2         | 0         |           |           |           |           |
| MEASUREMENT SCALE OF DEPENDENT VARIABLE                               |               |               |           |           |           |           |           |           |
| 15. nominal                                                           | 5             | 4             | 6         | 4         |           |           |           |           |
| 16. ordinal                                                           | 4             | 4             | 1         | 1         |           |           |           |           |
| 17. interval                                                          | 6             | 4             | 1         | 1         |           |           |           |           |
| NUMBER OF CATEGORIES OF INDEPENDENT VARIABLE                          |               |               |           |           |           |           |           |           |
| 18. two                                                               | 0             | 1             | 0         | 0         |           |           |           |           |
| 19. more than two                                                     | 0             | 2             | 1         | 0         |           |           |           |           |
| 20. post-hoc comparisons                                              | 0             | 2             | 0         | 2         |           |           |           |           |
| RELATIONS AMONG SAMPLES                                               |               |               |           |           |           |           |           |           |
| 21. independent samples                                               | 4             | 2             | 2         | 1         |           |           |           |           |
| 22. related samples                                                   | 2             | 1             | 2         | 3         |           |           |           |           |
| Themes                                      | Experts (n=6) | Novices (n=6) |
|-------------------------------------------|---------------|---------------|
| matched-pairs samples                     | 5             | 4             |
| repeated-measures samples                 | 5             | 1             |
| NUMBER OF GROUPS                           |               |               |
| one group                                  | 2             | 4             |
| two groups                                 | 6             | 4             |
| three groups                               | 3             | 0             |
| POPULATION PARAMETER                       |               |               |
| means                                      | 6             | 6             |
| correlations                               | 2             | 5             |
| medians                                    | 0             | 2             |
| regression coefficient                     | 0             | 1             |
| proportions                                | 4             | 3             |
| ISSUES OF VALIDITY                         |               |               |
| internal validity                          | 4             | 0             |
| external validity                          | 3             | 0             |
| TYPES OF DESIGNS                           |               |               |
| experimental research                      | 2             | 1             |
| research using pre-test, post-test         | 1             | 0             |
| trend analysis/timeseries/runs designs     | 1             | 2             |
| research in which causality can be inferred| 1             | 1             |
| research with retrospective data            | 1             | 0             |
| complex research designs                   | 3             | 0             |

*Themes given in bold indicate the ones that were addressed by a majority of experts (n>=4) in at least one of the tasks, research scenarios or statistical techniques, and the themes in italics were addressed by a majority of the novice group.

Table 5. Themes observed in the constructs related to theoretical aspects and the number of subjects who addressed the themes.

| Required Assumptions                      | Experts (n=6) | Novices (n=6) |
|-------------------------------------------|---------------|---------------|
| 1. normality                              | 5             | 4             |
| 2. equality of group variances            | 0             | 2             |
| 3. bivariate normality                     | 0             | 3             |
| 4. independence of observations            | 1             | 3             |
| 5. data transformations to ensure normality| 5             | 0             |
Table 6. Procedural themes addressed in the constructs of subjects and the number of subjects who addressed the themes.

| Themes                                                                 | Experts (n=6) | Novices (n=6) |
|-----------------------------------------------------------------------|---------------|---------------|
|                                                                       | research      | statistical   |
|                                                                       | scenarios     | scenarios     |
| TYPE OF INFERENCE PROCEDURE                                           |               |               |
| 1. directionality of hypothesis testing (one-tailed, two-tailed)       | 1             | 1             |
| 2. nature of inference procedure (i.e., hypothesis testing, estimation, fitting models) | 1             | 4             |
| 3. constructing a confidence interval                                 | 2             | 0             |
| COMPTUTATIONAL ELEMENTS                                               |               |               |
| 4. significance level (alpha) given/needed                            | 0             | 3             |
| 5. kind of test statistic used (e.g., $X^2$, $t$, $\chi^2$, $F$)      | 0             | 1             |
| 6. range of the values test statistic can take (i.e., being between -1 and +1) | 2             | 3             |

| 6. covariance matrices having compound symmetry                        | 2             | 0             |
| 7. linearity of relationships among variables                         | 2             | 2             |
| 8. robustness to violations of assumptions                             | 0             | 3             |
| **9. parametric/nonparametric tests**                                  | 6             | 5             |
| **10. number of required assumptions**                                 | 0             | 1             |
| **SAMPLE SIZE**                                                       |               |               |
| 11. applications of CLT for large sample versions                     | 1             | 2             |
| 12. required sample size for needed power (and other power considerations) | 2             | 2             |
| **UNDERLYING THEORETICAL DISTRIBUTION**                               |               |               |
| 13. cell size ($n \geq 5$) in contingency tables                      | 2             | 1             |
| 14. binomial distribution                                              | 2             | 1             |
| 15. exact/discrete distributions                                       | 1             | 2             |
| 16. F distribution                                                     | 0             | 1             |
| 17. t distribution                                                     | 1             | 1             |
| 18. chi-square distribution                                            | 0             | 2             |
| 19. complexity of mathematical relationships among statistical parameters involved | 0             | 1             |
| 20. distribution free tests/situations                                 | 0             | 3             |

Table 6 of 34
7. need for a standar table to look up | 0 | 0 | 0 | 2
8. method of determining degrees of freedom | 0 | 1 | 0 | 1
9. need for transformation of test statistics | 0 | 1 | 0 | 0
(i.e., Fisher's z transformation for \( r \))
10. consideration of outlier values during computation | 2 | 1 | 1 | 1
11. availability of population variance for computations | 0 | 0 | 0 | 1
12. data being usually organized in tables | 0 | 0 | 0 | 1
13. using row and column totals in calculations | 0 | 0 | 0 | 1
14. computing "pooled variances" | 0 | 1 | 0 | 0
15. mathematical calculations being relatively complicated | 0 | 1 | 1 | 3

### MEANING OF COMPUTATIONAL PROCEDURES

16. comparing within and between group variability | 0 | 1 | 0 | 1
17. comparing observed and expected values | 0 | 1 | 0 | 1
18. relating the order of one distribution with another distribution | 0 | 2 | 0 | 1

---

**Table 7. Non-technical themes observed in the constructs of subjects and the number of subjects who addressed the themes.**

| NON-TECHNICAL THEMES ABOUT STATISTICAL TECHNIQUES | Experts (n=6) | Novices (n=6) |
|---------------------------------------------------|--------------|---------------|
| REQUIRED ASSUMPTIONS                              | research     | statistical   |
| 1. popularity of the use of test                  | n/a | 0 | n/a | 3 |
| 2. class in which the technique is covered       | n/a | 0 | n/a | 4 |
| (Stat I, II, or III)                              |              |               |
| 3. level of conceptual difficulty of the technique| n/a | 0 | n/a | 1 |
| 4. how the test is named (after a person, by the  | n/a | 1 | n/a | 0 |
| underlying probability distribution, etc.)        |              |               |
| 5. degree of usefulness of the statistical technique | n/a | 0 | n/a | 4 |
| 6. statistical techniques being related to other  | n/a | 1 | n/a | 3 |
| tests (without specifying reasons)                |              |               |
| 7. the need for a textbook to remember the        | n/a | 0 | n/a | 2 |
| specifics of the statistical technique             |              |               |
| 8. generality of statistical techniques in terms   | n/a | 0 | n/a | 1 |
| of the situations they can be used for             |              |               |
| NON-TECHNICAL THEMES ABOUT                        |              |               |
| STATISTICAL TECHNIQUES                             |              |               |
| scenarios                                         |              |               |
| scenarios                                         |              |               |
| scenarios                                         |              |               |
| scenarios                                         |              |               |
| RESEARCH SCENARIOS                                                                 | 6   | n/a | 5   | n/a |
|----------------------------------------------------------------------------------|-----|-----|-----|-----|
| 9. kind of statistical method that the scenario suggests for analysis             | 3   | n/a | 3   | n/a |
| 10. evaluation of the research scenario for its general quality                  | 0   | n/a | 3   | n/a |
| 11. clarity of the design of a research scenario                                 | 3   | n/a | 4   | n/a |
| 12. factual observations from a research scenario                               |     |     |     |     |
| (e.g., age range of subjects, context of research, data being based on self-report, etc.) |

Cengiz Alacaci  
Department of Curriculum and Instruction  
Florida International University  
University Parck Campus  
Miami, FL 33199  
U. S. A.  
alacaci@fiu.edu

**Volume 12 (2004) | Archive | Index | Data Archive | Information Service | Editorial Board | Guidelines for Authors | Guidelines for Data Contributors | Home Page | Contact JSE | ASA Publications**