This is to certify that the

thesis entitled

MAKING INFERENCES FROM STATISTICAL SIGNIFICANCE TESTS:

A STUDY OF GRADUATE STUDENTS IN PSYCHOLOGY

presented by

Frederick W. Silver

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Psychology

Major professor

Date July 30, 1980

MAKING INFERENCES FROM STATISTICAL SIGNIFICANCE TESTS: A STUDY OF GRADUATE STUDENTS IN PSYCHOLOGY

By

Frederick W. Silver

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirement
for the degree of

DOCTOR OF PHILOSOPHY

Department of Psychology

1980

G 116135

ABSTRACT

MAKING INFERENCES FROM STATISTICAL SIGNIFICANCE TESTS:
A STUDY OF GRADUATE STUDENTS IN PSYCHOLOGY

By

Frederick W. Silver

Mounting criticism of statistical significance tests in psychology has raised the question of how well psychologists understand the limitations of this methodology. To answer this question, the present study investigated the types of inferences psychologists believe are valid based on statistically significant results from a single experiment, and whether psychologists understand the relationship between sample size, the p value, and strength of association. The study also assesses a number of educational and attitudinal variables, which were thought to be relevant to individual differences in beliefs and understanding.

A questionnaire entitled "Conceptions of Statistics" was specially developed. It consists of three main parts. The first part includes questions on educational background and career interests; coursework in statistics, research methods, the physical sciences and mathematics, and philosophy; and ten questions designed to assess attitudes towards psychologists as researchers, psychology as a science, and significance tests. The second part included two problems each presenting a synopsis of an experiment in psychology along with its statistical results. Each synopsis was followed by 36 possible

conclusions based on a completely crossed set of four factors: Tentativity Factor, Theoretical Generality Factor, Population Generality Factor, and Task Generality Factor. The same factorial structure was used for conclusions in both problems, though the specific content was different for each. Subjects were asked to judge whether each conclusion was valid or invalid.

The third part of the questionnaire contained four problems in which subjects were asked to compare two statistical results, each of which included a sample size, a statistic value, and a significance level. Sample size and p values were systematically varied. Subjects were asked to select the result which would give them greater personal assurance that the null hypothesis was false, or the result which manifested a stronger association between independent and dependent variables. Half of the problems used t-test scores and the other half correlation coefficients.

A sample of 80 graduate students in psychology at Michigan State University was randomly selected from the 160 students enrolled for courses in Spring term, 1979.

Responses to the attitude questions showed that a large number of subjects agreed that psychologists do shoddy research and disagreed that significance tests provide objectivity.

Pattern analyses were used to classify responses to the problems, and typologies were developed from the pattern analyses. These showed considerable individual differences in responses to the problems. A moderately strong tendency for subjects to generalize widely on Population, Task and Theoretical Generality Factors was observed. However, the four generalization factors were uncorrelated.

No one group of subjects consistently generalized more or less than others. Also, a majority of subjects made errors in judgments of personal assurance that the null hypothesis was false and strength of association. It appeared that many did not understand the relationship between sample size, the p value, and strength of association.

Some of the educational, career interest, and coursework variables were associated with the generalizations and errors, but no systematic pattern of relationships was observed. None of the attitude questions were related to generalizations and errors.

For the first two parallel problems, a comparison of pattern analyses, typologies and generalizations showed there was a marked inconsistency across problems. This was partially attributed to a random error process and partially to perceived differences in content. The randomness implied that the true associations between the educational, coursework, and attitude variables, and the generalizations and errors, were larger than what was observed.

Finally, some explanations for the individual differences in generalizations and errors were given, some implications for graduate education were noted, and some directions for future research were suggested.

To Josie,
for her steadfast support and patience,
with love and gratitude.

In memory of my father,
whose own career struggles
were a source of motivation and direction
for me.

ACKNOWLEDGEMENTS

I would like to thank the members of my dissertation committee for their help and support:

--John E. Hunter, who served as co-chair, for helping me consolidate my knowledge of statistical inference and generalization; for his presentation of an alternate point of view about psychological research; and for his help in the statistical analysis of the data.

--Griffith O. Freed, who served as co-chair, for his serious support of a clinician's interest in the basic questions and issues in pscyhology as a science; for his openness to new ideas; and for his time and energy.

--Norman Abeles and Raymond Frankmann, for their interest and suggestions.

I would also like to thank members of my family for their support and help during my long graduate career. And finally I would like to thank several faculty members of the Department of Psychology and Counseling Center who took a special interest in my intellectual and personal growth during my graduate studies: Al Rabin, Al Aniskiewicz, Bill Mueller, Dave Wenger and Joanne Hamachek.

TABLE OF CONTENTS

LIST OF TABLES

I. INTRODUCTION

The statistical significance test methodology was introduced over 50 years ago to help scientists overcome the uncertainty involved in making inductive inferences. This uncertainty arises from the desire to make generalizations across time and space, based on a limited amount of data from a limited number of cases. Psychologists have adopted this methodology and it has since become deeply ingrained in our way of doing science.

According to some observers the significance test methodology has become so institutionalized in psychology that it has come to function largely as a form of ritual in resolving the uncertainties of scientific generalization. Critics of the methodology maintain that very few understand its real limitations. However, others suggest that psychologists know but are inattentive to these limitations and intuitively compensate by using their own assumptions in making generalizations. Alternatively, it has been suggested that psychologists have ignored the mounting criticism of this methodology in part because of their belief that there is no other equally adequate and objective means of evaluating hypotheses, or analyzing data.

To what extent are psychologists cognizant of the limitations of the significance test methodology? How well do they understand it? The present study attempts to provide some data to begin to address these basic questions of what inferences and generalizations psychologists make using statistical significance tests.

II. REVIEW OF THE LITERATURE

Significance Test Methodology

The purpose of the significance test is to aid in making inferences about some characteristic of a population when only a sample of it can be observed. Significance tests are used in almost all research studies in psychology because it is rarely feasible to observe an entire population.

A simple prototype of research design in psychology is the experiment where two samples, or groups of subjects are observed. Subjects are randomly drawn from the population at large, and randomly assigned to the two experimental conditions. The two groups of subjects receive different experimental conditions and are then observed in terms of some dependent variable. Using the data from these observations various statistics are calculated for each of the two groups. Usually these are the means and standard deviations.

Suppose the two means of the dependent variable are different from each other. Should the experimenter then conclude that such a difference is characteristic of the entire population? In other words, should the experimenter conclude that it was the difference in treatment conditions that caused the difference in means and not that the two samples differed prior to the experiment in some way as a result of the natural variation expected with random sampling procedures (i.e., as a result of sampling error)? It is certainly possible that in the random selection and assignment process two experimental groups were

formed that differed on some important dimension, either from each other, or from the population at large; and that it was this difference, and not the difference in experimental conditions, that was responsible for the difference in sample means.

If the experimenter could determine whether the difference in means obtained in the two samples is the same as the difference that would be obtained if the entire population had been used in the experiment, there would be no question about sampling error and no need for a significance test. But this can never be known for sure without actually testing a very large sample (so that sampling error would be trivial).

Lacking this information, the experimenter must in some way take into account the possibility of sampling error in assessing the magnitude of the differences in sample means. This is where the significance test is used.

A significance test involves setting up a sampling distribution, a hypothetical relative frequency distribution, conditional on the assumption of the null hypothesis. This is the null hypothesis significance test.

In most cases the null hypothesis states that there are no differences in means (or other statistics). The alternate hypothesis is either the opposite of the null hypothesis, that there are differences in means, or in the case of the directional test (or one-tailed test), that mean X is greater than mean Y.

In setting up the sampling distribution the experimenter asks the question: If in fact there were no differences between population means, what would be the relative frequency of a difference of this

magnitude or more? This relative frequency is the p value. It is the probability that such a difference between means would occur if the null hypothesis were true (Type I error). If this probability is 5 percent or less, then by convention the null hypothesis is rejected and the alternate hypothesis is accepted.

While the significance test is usually described as an aid in making inferences from a sample to a population, it can also be thought of as a test against the competing hypothesis that sampling error (and not some real effect) is the cause of the difference obtained between groups. Thus, the significance test has been described as a test against the null hypothesis as a competing hypothesis (Carver, 1978). When a significance test yields a very small p value, then according to this framework, it can be said that the null hypothesis (of sampling error as the cause of the obtained difference) is not likely to be the correct explanation for the data.

The Fisher and Neyman-Pearson-Wald Models

In the history of the statistical significance test, two basic models and points of view have been advanced. The first of these was put forth by R. A. Fisher.

Fisher believed that the significance test provides a calculus for making provisional inductive inferences about the falsity of the null hypothesis. In Fisher's model, the experimenter formulates a null hypothesis and calculates a p value based on a hypothetical sampling distribution and the results of the experiment (Hogben, 1957). The experimenter is then in a position to reject the null hypothesis if its probability of giving the results obtained is very low. Fisher did not

necessarily advocate the choosing of a rejection region prior to the experiment, though he did generally recommend an .05 level. If an experimental result was obtained that did not meet this criterion then Fisher considered the result of the experiment to be inconclusive vis a vis the null hypothesis (Hogben, 1957). In Fisher's (1949) own words:

> It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result...It is usual and convenient for experimenters to take 5 per cent as the standard level of significance...(p. 13).
>
> ...it should be noted that the null hypothesis is never proved or established, but possibly disproved, in the course of experimentation (p. 16).

The main reason given by Fisher for never accepting the null hypothesis was because no probability of error could be attached to such an inference without assuming some specific rival hypothesis to be true (Fisher, 1955).

Fisher (1949; 1955) believed that significance tests provide a basis for a rigorous form of induction because they give a rigorous expression to the degree of uncertainty attached to a hypothesis. He noted that an inductive inference was always a provisional one, and could be enlarged or modified on the basis of new data.

E. S. Pearson, Neyman, and Wald developed a somewhat different statistical procedure that was called "decision tests." The decision test procedure provides a decision and risk calculus for making decisions under conditions of uncertainty and risk. Testing a hypothesis is viewed as a special case of the general decision problem (Wald, 1950).

The procedure of the decision test model, with regard to a

scientific hypothesis, is to set up a decision rule which specifies acceptance and rejection regions and which also specifies the consequences of such a decision rule. In the decision test model a decision can be made to either accept or reject the null hypothesis. This raises the issues of Type II error and power, whose development is associated with this group of statisticians (Bakan, 1970).

Contemporary significance test methodology is an amalgam of the significance test and the decision test models.

Problems in the Use of Significance Tests

Observations of the use of significance tests suggest that many psychologists do not fully understand this methodology and that auxiliary assumptions are used without ever being made explicit. Some have observed that we use significance tests in a very automatic way, as a "computational ritual" (Bakan, 1970; Hunter, 1979). According to Hazelett,

> We are so used to using our intuitions to fill in gaps in logic that we fail to notice that the model we are supposedly following cannot logically account for what we are actually doing. (1975, p. 61)

On the other hand, it is also possible that we are aware of these gaps but do not know of any alternative ways of evaluating data.

Some Common Misinterpretations of the p Value

Observations by Bakan (1970) and others suggest that there is much confusion about the meaning of the p value. Five common misinterpretations have been described in the literature.

(1) Confusing a conditional probability with an absolute probability.

The p value tells us the relative frequency a difference of the magnitude observed in the experiment would be obtained, if the null hypothesis were true and the experiment done an infinite number of times. Thus, the p value is a conditional probability based on the assumption that the null hypothesis is true. It does not provide a measure of the absolute probability that the null hypothesis is actually true. This absolute probability is unknown.

The failure to recognize that the p value is a conditional probability leads to the popular misconception that the p values gives the probability that the results of the experiment are due to chance (Carver, 1978), or that of believing that there is (1 - p) probability that the alternate hypothesis is true (see Wilson, 1961).

As Carver (1978) has pointed out, the p value cannot represent the probability that the results of an experiment were due to or caused by chance because the p value is calculated by assuming this to already be the case.

(2) Confusing the theoretical with the statistical hypothesis.

The p value as a conditional probability applies only to the null hypothesis being tested, not to the theory whose predictions were the basis for the experiment in the first place. Acceptance of the alternate hypothesis does not necessarily imply the validity of the theory. A common misinterpretation of the p value is to generalize its application to the theoretical hypothesis and to conclude that the probability that the theory has been confirmed is (1 - p). As Bakan

(1970) has noted, once an inference is made from the sample to the population (e.g., once the null hypothesis is rejected), then a secondary inductive inference is still required to confirm the theory. Usually, several experiments are needed to rule out rival theories. Thus, the results of a significance test may be said to be consistent or inconsistent with a particular theory, but this is not something that can be automatically inferred from a p value.

(3) Confusing inferences to the general with inferences to the aggregate.

The significance test was only designed to make inferences to the population as a whole (aggregate) from statistics computed from the sample as a whole (Bakan, 1970). Thus, it might be reasonable to tentatively assert on the basis of a very low p value that there would be differences between two means in the population as a whole. It would not, however, be valid to say that such differences exist for all members of the population, or all members of the sample for that matter. The type of statistics used in significance tests may be descriptive of the group as a whole, but not necessarily of every subject in the group.

An example taken from Bakan (1970) may be instructive in showing how insidious a confusion that is. Bakan describes what could be conceived of as a general experimental prototype in which twenty schizophrenics are compared with twenty normals on a dependent variable. Given this example, is it reasonable to assert on the basis of a very small p value that "schizophrenics differ from normals in such and such ways?" (p. 244). No, because this implies that all schizophrenics differ from all normals in such and such ways, and not that just the group of

schizophrenics differes as a group from the group of normals. According to Bakan, the p value obtained in such a study "bears only on the means of the populations and is not a 'measure' of the confidence that he (the experimenter) may have in his hypothesis concerning the nature of schizophrenia" (p. 245).

General and/or theoretical statements cannot be justified only by a significance test. They require secondary and tertiary inferences beyond that made from a sample to the population (Bakan, 1970). Many experimenters who do this are probably unaware that they are making additional inferences or auxiliary assumptions. These are rarely made explicit and rarely justified. Often there are no data available on which to base these assumptions however commonsensical or trivial they may seem.

(4) Using the p value as an automatic measure of the meaningfulness of observed differences.

As Bakan (1970) has observed, the p value is often construed as a "measure" of degree of significance, as an answer to the question: How significant are the results? This is manifested in the frequent practice of listing the p value for each significance test done alongside the statistical results. Or, in many cases asterisks are used to denote different degrees of significance. Thus, a p value of .01 (denoted by two asterisks) is seen as "better" than a p value of .05 (denoted by one asterisk), without any consideration of the strength of association between independent and dependent variables, and without regard for the type of relationship actually predicted by the substantive theory (e.g., one would expect a weaker association between variables whose

causal relationship is indirect).

Using the p value as an automatic measure of meaningfulness is also seen in the use of a significance test as a way of differentiating trivial versus non-trivial differences. Again, this is done in automatic ways, without regard to strength of association and the predictions of the substantive theory. This is seen in the practice of not reporting, or listing as "n.s.", statistics which do not reach conventional significance levels.

(5) Confusion about the relationship between N and p (the large N fallacy); confusion about the relationship between N, p and strength of association.

Similar to the interpretation of the p value as a measure of meaningfulness is the belief that a smaller p value represents a stronger effect or association between independent and dependent variables regardless of sample size. While it is generally true that the smaller p is, the greater the effect or association, p is a deceptive index of size of effect or association if sample size is not taken into account. This confusion about sample size, p, and strength of association, can be seen in the failure to understand that it takes a greater effect to yield "significant" results with a small sample than with a large one (Bakan, 1970). Conversely, it can be seen in the failure to understand that in a very large sample the size of effect necessary to generate "significant" results can be very small (e.g., with N = 400, it takes a relationship such as r = .10 to reach significance at the .05 level).

Underlying this misinterpretation about N, p, and strength of association is a confusion about the relationship of sample size to

p. Most psychologists have some awareness of the fact that if the null hypothesis is false it is easier to attain significance with a larger sample size, but few seem to really understand the fact that p is inversely related to N, independent of the deviation from the null hypothesis (Bakan, 1970). Thus, for any given deviation from the null hypothesis (e.g., of no differences between means), the larger the sample size, the greater the probability of rejecting the null hypothesis (obtaining a "significant" p value), when it is false.

A study by Rosenthal and Gaito (1963) bears on this confusion about the relationship of sample size and p. In this study nine faculty and ten graduate students of the Department of Psychology of the University of North Dakota were asked to rate their degree of belief or confidence in hypothetical significance test results with a variety of p levels, at sample sizes of ten and 100. The investigators found that for equal p values both faculty and graduate students expressed more confidence when the sample size was 100 in contrast to ten. They interpreted this to mean that decreased probability of Type II error (because of increased power) was taken into account in the greater confidence given the larger sample size. Their hypothesis could be rephrased as: subjects confused probability of Type II error, which decreases with larger N, with probability of Type I error (the p value), which was given as constant by the experimenters.

Bakan (1970) offers a different interpretation of the Rosenthal and Gaito finding. He believes that the results reflect a misunderstanding of the relationship between sample size and p on the part of subjects. Bakan argues that a smaller N reflects a relatively greater

effect than a larger N, for any given p value, "and if the p value is the same the probability of committing a Type I error remains the same. Thus one can be more confident with a small N than a large N" (1970, p. 241). Bakan's conclusion--that one can have greater confidence with a smaller N--does not follow from his premise--that if p values are equal then regardless of N probability of Type I error is constant. For if confidence is defined as confidence that the null hypothesis is false then the p value alone is a rational criteria on which to make judgements, and N is irrelevant (actually N is already taken into account in looking up p). It seems, then, that Bakan is either confusing confidence with strength of association, or simply defining it as such.

Bakan suggests that the confusion about the relationship between N and p, the large N fallacy, reflects two errors: one, that the p value is a "measure" of confidence; and two, the lack of understanding that the p value is a function of sample size. According to Bakan, the typical thinking behind this confusion goes like this:

> The p value is a measure of confidence; but a larger number of cases also increases confidence; therefore, for any given p value, the degree of confidence should be higher for the larger N (p. 241).

However, since Rosenthal and Gaito did not explicitly define confidence in their study it is hard to know why subjects expressed greater confidence in larger sample sizes.

Rosenthal and Gaito also reported in their study an exponential curve of increasing confidence across decreasing p values, with a somewhat precipitous increase in confidence occurring between p = .10 and p = .05. This was thought to reflect the scientific convention of considering .05 as the demarcation between acceptance and rejection of

the null hypothesis.

General Problems in the Use of Significance Tests

(1) Meehl's paradox: the significance test and theory corroboration

There are a number of critics of the significance test methodology who question the rationality of the entire methodology as well as its usefulness as a way of doing science (Hazelett, 1975; Meehl, 1978). Meehl, one of the most outspoken of these critics, has recently written:

> I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas (of psychology) is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology (1978, p. 817).

To understand Meehl's criticism, it is helpful to be familiar with the approach to science recommended by Karl Popper (1959; 1963).

Popper has recommended that scientists attempt to corroborate theories by subjecting them to a grave risk of refutation; the greater the risk of refutation, the greater the degree of corroboration. Popper's recommendation is based on the fact that refutation of a theory, following the failure of its predictions to hold true, uses the valid form of implication known as *modus tollens* (denying the consequent). However, confirmation of a theory, following the success of its predictions, uses an invalid form of implication (the fallacy of affirming the consequent). It is from this logical asymmetry that Popper has derived his recommendation that scientists strive to refute rather than

confirm theories.

According to Popper, the risk a theory is subjected to in an attempt to falsify it is a function of the range of outcomes specified as inconsistent with the theory. The greater this range of forbidden outcomes, the greater the degree of corroboration when these do not occur. Thus, Popper believes that scientists should strive for highly "informative" theories, which provide very clear and specific predictions.

Meehl's paradox is the following:

> In the physical sciences, the usual result of an improvement in experimental design, instrumentation, or numerical mass of data is to increase the difficulty of the "observational hurdle" which the physical theory of interest must successfully surmount; whereas, in psychology and some of the allied behavioral sciences, the usual effect of such improvement in experimental precision is to provide an easier hurdle for the theory to surmount. Hence what we would normally think of as improvements in our experimental method tend (when predictions materialize) to yield stronger corroboration of the theory in physics, since to remain unrefuted the theory must have survived a more difficult test; by contrast, such experimental improvement in psychology typically results in a weaker corroboration of the theory, since it has now been required to survive a more lenient test (1970; p. 252-253).

Meehl's paradox is a result of the use of the null hypothesis significance test in psychology. The paradox is derived from the belief that the point-null hypothesis is almost always false in psychology. This belief is supported by observations that in multiple significance tests with large samples of subjects, very high percentages achieve significance (Nunnally, 1960; Meehl, 1970; Bakan, 1970; Berkson, 1970). For example, Nunnally reports that in a sample of 700 subjects in a public opinion study nearly all the correlations with individual

difference variables were significant. According to Nunnally, "if the null hypothesis is not rejected it usually is because the N is too small. If enough data is gathered, the hypothesis will generally be rejected (p. 643)."

Another example which supports the belief that the null hypothesis is almost always false in nature comes from Bakan (1970). Bakan reports that data from 60,000 subjects from all over the United States proved significant for every significance test run, even for tests that were set up by dividing subjects into "arbitrarily" selected groups, such as those east of the Mississippi versus those west of the Mississippi. Tests on a population of this size, noted Bakan, were highly significant with only minute differences in groups.

There is, however, at least one published report of an experiment in the behavioral sciences with a very large N in which significance tests did not yield significant results. Oakes (1975) notes that reports of the U.S. Office of Economic Opportunity of 23,000 subjects, 10,000 of whom were randomly assigned to a special educational program, showed no significant differences in achievement between groups. Based on this experiment as a counterexample to Meehl, Oakes suggests that it is only in the self-selected groups (e.g., nonrandomized) research design in psychology that the null hypothesis always is false in nature. Meehl (1978), however, disagrees. He believes the null hypothesis is false in all types of research designs.

The explanation provided by Meehl (1970; 1978) for why the null hypothesis is always false in research using a nonrandomized design is that <u>all</u> variables measuring individual psychological

characteristics (e.g., personality traits, values, aptitudes, demographic characteristics, etc.) are in some way causally related, however indirectly and however weakly. Thus, one would never obtain a measure of association between two such variables that was zero. The explanation given for the falsity of the null hypothesis in randomized, true experimental designs is that there are many small situational or artifactual influences which are elicited in any experimental group which cause an effect that will, in a large enough sample, generate significant results:

> It would require considerable ingenuity to concoct experimental manipulations...where one could have confidence that the manipulation would be utterly without effect upon the subject's motivational level, attention, arousal, fear of failure, achievement drive, desire to please the experimenter, distraction, social fear, etc....Suffice it to say that there are very good reasons for expecting at least some slight influence of almost any experimental manipulation which would differ sufficiently in its form and content from the manipulation imposed upon a control group...(Meehl, 1970, p. 260).

Meehl observes that the substantive theory usually makes a directional prediction. He asserts that in an experiment with perfect power, it is possible for the outcome of a directional significance test to be either "successful" (confirms the theory) or "unsuccessful" (disconfirms the theory), irrespective of the truth of the substantive theory. That is, since there is almost always a non-zero difference in the dependent variable (even for a theory with no truth value whatsoever), in some cases the group designated as the "experimental" group will have a higher mean, and in other cases the group designated as the "control" group will have the higher mean. If a theory has no connection with the truth, there will be no relation between the direction of difference predicted by that theory and the direction of artifactual

difference obtained. And thus, whether the theory's predictions are correct or incorrect will be strictly random. Stated in other words, in about half the cases a theory (with no truth value) will correctly predict the direction of the artifactual effect, assuming all the experiments are at perfect power. Thus, for a series of experiments at perfect power there is a prior probability of .50 that the null hypothesis significance test will be statistically significant in the direction of confirming the theory.

According to Meehl, then, for a theory that has no truth value whatsoever, the expected relative frequency of "pseudosuccessful" outcomes of a directional significance test at perfect power will be approximately .50. And so as experimental precision (power) increases, the expected relative frequency (and prior probability) of "pseudosuccessful" significance test outcomes approaches .50.

Meehl (1970) concludes,

> that the effect of increased precision, whether achieved by improved instrumentation and control, greater sensitivity in the logical structure of the experiment, or increasing the number of observations, is to yield a probability approaching ½ of corroborating our substantive theory by a significance test, <u>even if the theory is totally without merit</u>...It goes without saying that successfully negotiating an experimental hurdle of this sort can constitute only an extremely weak corroboration of any substantive theory...(p. 262).

In short, as experimental precision increases, the prior probability of "pseudosuccessful" outcomes increases, hence the risk of refuting a theory decreases, diminishing the potential degree of corroboration of that theory. This is, according to Meehl, exactly the opposite of what occurs in physics and other sciences, which set up the theoretical hypothesis, not a null hypothesis, as the test hypothesis in any

experiment.

In order to remedy this situation Meehl (1978) advocates abandoning significance tests in favor of a methodology in which psychologists test theories by making more specific quantitative predictions, e.g., point predictions, interval predictions, and if these cannot be derived from the theory, predictions about the form of the function, or at least, order of numerical values or numerical differences.

The major problem with Meehl's paradox is that it assumes that a "successful" significance test is automatically and simple-mindedly taken by psychologists as corroboration of the theory, without regard for the size of effect, or strength of association between independent and dependent variables. This may, in fact, occur only rarely, in which case the paradox would have little relevance to actual scientific practice. However, while the paradox may be an empty one, Meehl's suggestions for a more quantitative approach to theory testing provide a valuable alternative to the significance test.

(2) Strength of association and quantitative data evaluation.

Another problem with the null hypothesis significance test is that it does not provide information about strength of association between variables. As noted above, the problem with the using of the p value to index strength of association is that p is very much affected by sample size and is thus a poor, if not deceptive, index. Also, in a general way, the significance test methodology fails to provide and probably even discourages, a quantitative approach to data evaluation. A quantitative approach is where the data are evaluated to determine

the strength and pattern of functional relationships between variables, rather than to see if a particular relationship exists or doesn't exist (which is what the significance test, at best, assesses).

(3) The problem of sampling error.

Hunter (1979) has recently observed that most psychologists falsely believe that the significance test solves the problem of sampling error. They believe that a significance test can eliminate the uncertainty in empirical results from any one study, and thus that a well-designed study can be truly definitive.

The misunderstanding of significance tests and sampling error is also responsible for the misleading procedures used in reviewing and integrating areas of research (Hunter, 1979). The practice Hunter is critical of is where a reviewer tallies both the significant and the nonsignificant findings concerning some phenomenon and then ascribes the almost inevitable discrepancy that results either to methodological problems in one or more studies, or takes this discrepancy at face value (as real); in which case, the reviewer then offers suggestions for research that would help to tease out the very subtle interactions responsible for the more superficial conflict in results. It is these misleading review practices, according to Hunter--based as they are on a misconception of significance tests--which are responsible for the common perception among psychologists that our body of knowledge is riddled with discrepancy and lack of consensus.

Hunter advocates the use of confidence intervals as a substitute for significance tests in the analysis of data in an individual study. He suggests that the use of confidence intervals would make obvious to

researchers the extent of uncertainty that really exists in their data as a result of sampling error. He further argues that the only real way to eliminate this uncertainty is to do research with very large samples, or since this is rarely feasible, to cumulate findings across similar studies.

Some have argued that the significance test has no place in psychology and that alternate ways of hypothesis testing and data evaluation need to be adopted (Meehl, 1978; Hazelett, 1975; Hunter, 1979). Others, however, take the more moderate position that when used properly the test provides some useful information in the always uncertain business of scientific inference (Bakan, 1970).

The criticism of the significance test methodology has been mounting for over a decade. To what extent have psychologists taken notice? To what extent are they aware of the problems and limitations of this methodology? These are questions that have not yet been investigated empirically.

The Present Study

The present study was undertaken to investigate psychologists' beliefs about the validity of different inferences which might be made using statistical significance tests. Also, it was designed with the goal of investigating their understanding of some quantitative and statistical concepts that are involved in this methodology. Finally, the study was designed to assess factors which might help explain individual differences in beliefs about validity and in understanding of statistical concepts.

Specifically, the study investigates the types of inferences

graduate students in psychology believe are valid based on a single experiment with statistically significant results. In effect, then, it investigates the degree to which a subject is willing to generalize beyond a particular experimental situation. The study also investigates the understanding of concepts such as strength of association and personal confidence that the null hypothesis is false, given different significance test results and sample sizes. Finally, it examines a number of biographical, educational and attitudinal variables thought to be relevant to beliefs and understanding.

III. METHOD

The Instrument

The instrument used for this study is a specially designed questionnaire entitled, "Conceptions of Statistics" (see Appendix A). It has been pilot tested and revised several times.

The questionnaire consists of three main sections. The first of these is "Biographical Information." The section on biographical information includes identifying information; questions on educational background, graduate level and program of study, professional interests, and coursework in areas such as statistics, research methods, natural sciences, mathematics, and philosophy. It includes attitude questions about psychology as a science and about the use of statistics in psychology. These attitudinal questions were developed to assess such qualities as naivete and optimism versus cynicism and skepticism about research in psychology, psychology as a science, and about the use of significance tests. These attitudinal questions, along with the biographical items on education and professional background, were used to help identify factors involved in the individual differences obtained in the second two sections of the questionnaire.

The second section of the questionnaire consists of Problem 1 and Problem 2. Problems 1 and 2 are identical in format. They both consist of a one page description of a research study and its statistical results, followed by 36 possible conclusions that might be drawn from

that study. For each of these 36 possible conclusions, subjects were asked to indicate whether it was valid or invalid based only on the design of the study and its statistical results. Subjects were also asked to indicate which of the 36 conclusions was the best.

Problem 1 and 2 were chosen and adapted from actual research studies in psychology. They were selected on the basis of five criteria: (1) of general interest; (2) non-controversial, so as to minimize biases; (3) reasonably simple in design; (4) typical of psychological research; and (5) designed to test some theory. Problem 1 is a study about the cognitive processes involved in the evaluation of syllogisms (based on research by Carroll, 1976). Problem 2 is a study about differences in the cognitive functioning of the two cerebral hemispheres (based on work by Kimura, 1969, and Wagner, 1976).

Problems 1 and 2 were designed to assess subjects' interpretation of significance tests in a concrete situation. The items following the research descriptions and statistical results were constructed in a completely crossed factorial design using four factors (3 X 3 X 2 X 2; see Table 26). The levels of these four factors can be said to differ in their degree of generalization.

Factor A was the Format or Tentativity Factor because its three levels presented conclusions in different formats in such a way as to vary the degree of tentativity attached to each. At the first level of this factor (Nontentative level) conclusions were presented without any indication of tentativity. Conclusions in levels two and three of this factor (Tentative and Tentative-Qualified levels) were both described as tentative. Level three differed from level

two by the addition of a final clause specifying for this conclusion its probability of error due to chance. This third level was included to assess whether subjects failed to recognize that the p value is a conditional probability (e.g., believed a p value could be taken to mean the probability of error due to chance).

Factor B, the Population Generality Factor, also had three levels of generalization. At the first level, or level of least generalization, conclusions were made with specific reference to the population from which the study's sample was hypothetically drawn. This level is hereafter called the Population Specific level. In level two, the College Sophomore level, all conclusions were described as applying to American college sophomores. In level three, the Population Unqualified or Population General level, no restrictions were given for the population to which the stated conclusion applied. The third level of the Population Generality Factor was used to assess whether subjects were willing to make inferences to the general as well as to the aggregate.

Factor C, the Task Generality Factor, had two levels. Conclusions were either written to apply to the specific type of task used in the experiment (level one), or without any qualification with regard to task (level two). These levels were called, respectively, the Task Specific level and the Task Unqualified level. Like Factor B, Factor C was used to assess the degree to which subjects were willing to generalize across time, space, and experimental conditions. Since the questionnaire instructions clearly indicated that judgements as to the validity of various conclusions were to be based only on what

was presented in the research descriptions and statistical results, Factors B and C provided information as to the degree to which subjects were willing to generalize from the results of an individual experiment.

Factor D was the Theoretical Generality Factor. Its two levels were the Operational level and the Theoretical level. In the Operational level conclusions were stated in terms of the operational hypothesis being tested in the study. In the Theoretical level conclusions were drawn concerning the substantive theory that hypothetically motivated the experiment in the first place. Factor D was designed to assess the degree to which subjects made inferences from the statistical to the theoretical hypothesis.

The third section of the questionnaire consisted of Problems 3 through 6. Problems 3 through 6 were based in part on the work of Rosenthal and Gaito (1963). They were an attempt to replicate and expand their findings. These three problems were used to assess subjects' understanding about the relationship between sample size, p value, and both confidence and strength of association. They were also used to determine if subjects differentiated between confidence and strength of association.

Problems 3 through 6 each consists of pairs of significance test results which selectively differed on N, t-test score or correlation coefficient size, and p value. There are two sets of pairs for each problem (e.g., each of the four problems had two main parts or subsections numbered I and II), with the N and p values for each set being constant across problem subsections.

In Problems 3 through 6 subjects were asked to make a set of

comparisons between two significant test results. Each significant test result included a value for N (a lower case "n" was actually used in the body of the questionnaire), a value for a statistic, either t or r, and a value for p, such that any comparison was in effect a comparison of these three parameters.

Problems 3 and 5 asked subjects to compare two significant test results at a time and judge which of the two would give them more personal assurance that their null hypothesis was false. The term personal assurance was used instead of confidence so as to avoid any technical connotations the latter might suggest. Problems 4 and 6 were identical to 3 and 5 except that instead of asking to judge personal assurance, subjects were asked to compare the strength of relationships between the independent and dependent variable as manifested in each of the pair of significance test results, and to indicate which was the stronger of the two.

The factorial design of Problems 3 through 6 is presented in Table 1.

Table 1

THE FACTORIAL STRUCTURE OF THE THIRD SECTION OF THE QUESTIONNAIRE

	JUDGEMENT TYPE	
STATISTIC TYPE	Personal Assurance	Strength of Relationship
t-test scores	Problem 3	Problem 4
correlations	Problem 5	Problem 6

In addition to asking for a choice of the particular significance

test result which would give greater personal assurance or which manifested a stronger relationship, subjects were asked to indicate why they made the choice they did. Specifically, subjects were asked to indicate which of three factors (the comparisons of three pairs of variables, N, t or r, and p) figured either positively, negatively, or not at all in their choice by checking one of five boxes for each possible factor. The five boxes allowed subjects to indicate whether that factor was the most important factor, second most important factor, third most important factor, not a factor, or was a negative factor in their choice.

Coding

Data from the questionnaire were coded and then keypunched on computer cards. Codes for open-ended variables were established by the experimenter after looking through all questionnaires for the item in question.

A special coding scheme had to be developed for the parts of Problems 3 through 6 in which subjects were asked to indicate why they had chosen the statistical result above which they did. This was made necessary because responses to this segment of Problems 3 through 6 had different meanings depending on which choice the subject had made. Thus, the special coding was set up to take into account the choice of statistical result for that problem.

The scheme developed for this purpose used a weighted (by factor importance) directional preference score. This score seemed like a straightforward way of representing both degree of importance and direction of preference (e.g., for high or low values) for each of

the three parameters or factors (N, t or r, and p) that subjects could use in making their decisions.

For each comparison of two statistical results a score between -3 and 3 was determined. Negative scores were used when subjects endorsed lower values of N, t or r, and higher values of p. Positive scores were given for endorsement of higher values of N, t or r, and lower values of p. Direction of preference for each comparison was determined by the particular statistical result chosen, whether the values for the three parameters of that choice were higher or lower than the corresponding parameters for the other result, and whether the parameter preferred was described as a positive or negative factor in the subject's choice of statistical result.

Weights from 0 to 3 were determined for each parameter by the importance of that parameter in the subject's choice. A weight of 3 was assigned if the parameter was checked as the most important in the subject's choice. A weight of 2 was assigned if the parameter was checked as second most important. A weight of 1 was assigned if the parameter was third most important. A weight of 0 was assigned if the parameter was not a factor at all. A weight of -1.5 was assigned if the parameter was checked as being a negative factor. (A weight of -1.5 attached to a negative directional preference yielded a positive directional score, specifically a score of +1.5.)

In coding the third section of the questionnaire it became evident that a large portion of subjects either misunderstood or didn't read carefully the instructions describing how to indicate which factors determined their choice of statistical results in the opening

part of each problem. These subjects failed to use when appropriate the fifth column for checking a negative factor in their decision. This failure was associated with repeated and predictable inconsistencies in directional preference across subsections of problems. These inconsistencies were clearly a function of the predetermined variation in parameter values (e.g., in some statistical results high N was purposefully associated with low $\underline{p}$, while in others it was associated with high $\underline{p}$) which for many decision strategies required that subjects use the negative factor column in order to maintain consistency. In some cases the failure to understand or follow instructions in these parts of Problems 3 through 6 led to subjects alternately preferring first low $\underline{p}$ then high $\underline{p}$, while maintaining a consistent preference for high r and for high N. In these cases it was eminently clear that subjects had failed to follow the questionnaire's written instructions.

For subjects who failed to use the negative factor column and who showed predictable inconsistencies in their directional preferences, a special recoding procedure was used to recode their scores. In this procedure the experimenter in effect simulated the scores that would have resulted had subjects been aware of and used correctly the negative factor column provided in this part of the questionnaire. In order to do this the subject's decision scheme had to be determined for all problems. This was usually easy because many of these subjects checked identical boxes for groups of problems. After the decision scheme was determined the experimenter, in effect, had to go through and check in those boxes the subject should have used. From this a recoded score could be determined according to the standard coding

procedures.

A special variable was added to the keypunched data to indicate whether subjects' responses for these segments of Problems 3 through 6 had to be recoded. This was used so that data analysis could be performed both with and without these subjects if need be, and to help in the assessment of individual differences.

Sampling and Administration Procedures

The sample for the present study was 80 graduate students in psychology at Michigan State University. In order to obtain this sample 102 students were randomly selected from the 160 graduate students in psychology that were registered for classes for Spring term, 1979. Two students had to be eliminated from this group because they had served as subjects in the pilot testing of the questionnaire.

All the remaining 100 students were then contacted by telephone by the experimenter, with the exception of three students who were living long distances from East Lansing. A fourth student was recuperating from surgery and was unable to participate in this study. Of the remaining 96 students who would be able to be subjects for the experiment 91 agreed to participate. The reason for not participating given by those who refused was lack of time, or in one case, disinterest. Of the 91 students who agreed to participate, 80 eventually came in and completed the questionnaire. Of these 80 subjects, 67 were able to come to one of the six group administration sessions; the remaining 13 had to be scheduled individually or in two-person groups.

When called about participating in the research project subjects

were all offered $2. With the exception of five students who declined to take the incentive, all subjects were paid in cash following completion of the questionnaire.

In addition to the instructions within the questionnaire, all subjects were told verbally by the experimenter to read the written instructions carefully, and to ask any questions they had concerning the questionnaire. On the whole administration went very smoothly and subjects only occasionally asked clarifying questions. Average questionnaire completion time was approximately 35-40 minutes.

Order Variations in Administration

In order to evaluate the possibility of two different sets of order effects, the third section of the questionnaire was administered in four different orders. Table 2 summarizes the two primary order factors and the 2 X 2 design that was used to evaluate them.

Table 2

ORDER VARIATIONS IN QUESTIONNAIRE ADMINISTRATION

	ORDER FACTOR I (JUDGMENT ORDER)	
ORDER FACTOR II (STATISTIC ORDER)	Assurance before Strength of Relationship	Strength of Relationship before Assurance
t's before r's	3, 4, 5, 6*	4, 3, 6, 5*
r's before t's	5, 6, 3, 4*	6, 5, 4, 3*

*Order of Problems 3 through 6 in the questionnaire booklet

Questionnaires with the four different orders were interleafed in a fixed order and as subjects arrived they were given the top

questionnaire. Thus, the four different ordered questionnaires were distributed to 20 subjects each on the basis of order of participation in the study.

The importance of the two order factors, Judgment Order and Statistic Order, was evaluated with an analysis of variance on the directional preference scores. Directional preference scores were used rather than choice responses because it was thought that they would be more sensitive to any potential order effects.

The design submitted for the analysis took the form: A x B x C x D x P x Q (S/PQ), where A was the Parameter Comparison Factor (N, statistic, p), B was the Subsection Factor (I, II), C was the Judgment Factor (personal assurance, strength of relationship), D was the Statistic Factor (t, r), P was the Judgment Order Factor, Q was the Statistic Order Factor (see Table 2), and S was the Subject Factor (subjects were nested in combinations of P and Q). Following the analysis of variance eta was calculated for each effect.

None of the three basic between groups effects--P, Q, and PQ--were statistically significant. The eta-squared's for these three effects are either tiny (.003 for PQ) or infinitesimal (less than .0000 for P and Q). Thus, at the between groups level none of the three effects are of any importance in accounting for the variance in the dependent variable. Of the higher-order interactions with the order factors, seven out of 48 effects were statistically significant at the .05 level--one out of 16 for P, four out of 16 for Q, and two out of 16 for PQ. The average eta-squared for these seven significant effects is .0017. Considering the very tiny size of these eta-squared's and the

fact that the seven significant higher-order effects were not supported by substantive (nontrivial) lower-order effects, it seems safe to conclude that order variations were of very small or negligible importance in the directional preference scores. For this reason they are ignored in further analyses of data from Problems 3 through 6.

No variation in order of administration was necessary for Problems 1 and 2. These two problems were not directly compared. They were two sources of data used to answer the same questions. Comparisons done on these two problems were done to see if subjects followed a methodological rule and responded consistently across problems or whether response rules were content-bound and differed with the different content in problems. Also, the experimenter was willing to assume there would not be any second-order interactions between order of presentation of Problems 1 and 2, and all the other variables in the questionnaire.

The Research Questions

Question 1: What are the background and coursework characteristics of this sample of graduate students in psychology?

Question 2: What are their attitudes on statistical significance tests, psychology and psychologists?

Question 3: How are these attitudes related to background and course-work characteristics?

Question 4: What kinds of research conclusions do subjects believe are valid in Problems 1 and 2? What rules or strategies can be inferred from these responses?

Question 5: What kinds of generalizations do subjects make in Problems 1 and 2? Are these related to the miscellaneous variables for each problem?

Question 6: To what extent are subjects consistent in response strategies and generalizations in Problems 1 and 2?

Question 7: What strategies do subjects use to make judgments about strength of relationship and personal assurance, for the two statistics presented in Problems 3 through 6?

Question 8: In what ways do the rules for judging personal assurance and strength of relationship differ from ideal or correct rules?

Question 9: What biographical characteristics (background, coursework, attitudes) are related to generalizations and errors in Problems 1 through 6?

IV. RESULTS

Question 1: What are the background and coursework characteristics of this sample of graduate students in psychology?

Background Variables

In order to provide some descriptive information about the graduate students in the sample, responses to the biographical section were tabulated and are described below.

The sample was almost equally divided between males and females: 49 percent males and 51 percent females. The mean age was 27.5 with a standard deviation of 4.33.

Graduate Education Characteristics

The most frequent major field of study was Clinical Psychology (49 percent), followed by Industrial-Organizational Psychology and Social-Personality Psychology, each with 13 percent, Developmental Psychology with 10 percent, Ecological Psychology with 9 percent, and Experimental Psychology with 8 percent. There were no students from the Quantitative Psychology program.

The most frequent minor field listed was none with 20 percent; this was followed by developmental psychology with 19 percent; clinical psychology (including special clinical areas such as grief and loss counseling) with 11 percent; social-personality psychology and industrial-organizational psychology (including related business or management

minor areas), both with 9 percent; physiological psychology, neuropsychology or biological science with 8 percent; quantitative and methodological psychology, ecological psychology, and other fields of social science, each with 6 percent; and experimental psychology with 1 percent.

The mean number of years of graduate study for the sample was 3, with 17 first year students, 15 second year students, 16 third year students, 15 fourth year students, 8 fifth year students, 5 sixth year students, 2 seventh year students, and 2 students who did not answer this item. Of the 77 students who answered the question concerning number of years of graduate work in other social or natural sciences, 71 had none. The remaining 6 subjects had anywhere from 1 to 5 years, with a relatively even spread across this range.

Among the background variables, the number of years of graduate work in psychology was directly related both to age ($r = .33$), and to the total number of graduate level statistics courses taken ($r = .43$). It was not related to major field (eta = .12, omega = .0), but it was related to minor field (eta = .53, omega = .42).[1] Number of years of graduate work in other social or natural sciences was also directly related to age ($r = .31$) as well as number of statistics courses taken outside the Department of Psychology ($r = .24$).

Undergraduate Major and Minor

The most common undergraduate major given was psychology with 76 percent; this was followed by a double major of psychology with

---

[1] Eta and omega are based on analysis of variance tables and are calculated using the following formulas (see Hays, 1963): Eta-squared = SS_{Bet}/SS_{Total}. Omega-squared = $(SS_{Bet} - (J-1)MS_{With})/(SS_{Total} - MS_{With})$

another social science, with 5 percent; another social science without psychology, with 5 percent; psychology with a humanity as a double major, 3 percent; a humanity without a double major in psychology, 3 percent; and education, business, and mathematics, each with 1 percent.

Over half the sample listed no undergraduate minor (51 percent). The next most frequent response was some area of the humanities (13 percent). This was followed by mathematics or computer science, with 9 percent; natural sciences and social sciences (other than psychology), each with 8 percent; psychology with 4 percent; a combined humanity with social or natural science, with 4 percent; education with 2 percent; and 2 percent which was coded as miscellaneous.

Career Interests

The frequencies of subjects' ranking of their career interests are presented in Table 3.

Table 3

FREQUENCIES OF RANKINGS OF CAREER INTERESTS FOR CLINICAL AND NON-CLINICAL MAJORS

Career Interest		Ranking				
		1	2	3	4	5
research	CL	2	6	9	12	10
	NCL	25	8	6	2	0
teaching	CL	1	10	18	6	4
	NCL	5	17	10	3	2
clinical	CL	35	1	2	1	0
	NCL	4	1	4	7	25
consult.	CL	1	21	6	10	1
	NCL	7	9	16	7	2
admin.	CL	1	0	4	10	24
	NCL	0	2	5	19	15

The most preferred professional interest was clinical work for clinical

(CL) students and research for non-clinical (NCL) students. CL and NCL students differed most clearly in the preference for clinical work, research, and teaching, the latter two activities finding greater preference among NCL students.

Math Ability

Subjects' perception of their own math ability varied widely, with 9 subjects rating themselves as poor, 18 as fair, 26 as good, 25 as very good, and 2 as excellent. Students who majored or minored in math as undergraduates perceived themselves as having greater ability in this area (point-biserial r = .35). They were also likely to have had more terms of math and physics (point-biserial r's = .57, .36). Also, there was a tendency for subjects with different graduate school minors to perceive their math ability differently (eta = .45, omega = .32). It was not surprising that the group of subjects with quantitative or methodology minors rated their math ability higher than did most other groups.

Coursework

The mean number of terms of statistics, research methods, math and physical sciences, and philosophy are presented in Table 4. In addition, 31 subjects (39 percent) indicated they had taken a philosophy course in logic; 23 (29 percent) had taken a course in the philosophy of science; and 20 (25 percent) had taken a course in epistemology.

Table 4

SUMMARY OF COURSEWORK IN STATISTICS, RESEARCH METHODS MATH AND PHYSICAL SCIENCES, AND PHILOSOPHY

Subject Area	Mean	S.D.
Graduate level terms statistics - Psych. Dept.	2.4	1.3
Graduate level terms statistics - Other Dept.	.8	1.2
Graduate level terms statistics - Total	3.0	1.7
Years since last statistics course	1.5	1.7
Undergraduate terms - statistics	2.2	1.4
Graduate level terms - research methods	1.5	1.5
Undergraduate terms - research methods	1.4	1.1
Number of terms - math	2.4	2.4
Number of terms - engineering	.3	1.3
Number of terms - physics	.8	1.3
Number of terms - chemistry	1.3	2.1
Number of terms - total sciences and math	4.9	6.1
Number of terms - philosophy	2.1	2.0

Coursework in Statistics

The total number of graduate level terms of statistics was related to major (eta = .55, omega = .50) and minor (eta = .60, omega = .52) fields of study. Subjects in the Ecological Psychology (ECO) graduate program had the highest average number of total statistics courses, followed by subjects in the Industrial-Organizational (I-O) program, the Social-Personality (S-P) program, the Developmental (DEV) program, the Clinical Program (CL), and the Experimental (EXP) program. In general, subjects in the CL program had a smaller number of statistics courses than subjects in the NCL programs (point-biserial r = -.33).

The total number of graduate level terms of statistics was broken down into number of terms taken in the Department of Psychology and number of terms taken in other departments. Number of terms of

statistics taken in the Department of Psychology was related both to major (eta = .48, omega = .42) and minor (eta = .50, omega = .39) field of study. In descending order, the major fields ranked as follows: I-O, CL, S-P, DEV, ECO, and EXP. In this case there was no difference between CL and NCL subjects (point-biserial r = .07).

Number of terms of statistics taken in another department was also related to both major (eta = .63, omega = .59) and minor (eta = .56, omega = .47) fields of study. In descending order the major fields ranked as follows: ECO, DEV, S-P, I-O, EXP, and CL. In general, CL subjects had a smaller number of terms than NCL subjects (point-biserial r = -.44).

As to be expected, the number of years since the last statistics course was directly related to age (r = .35) and to number of years of graduate study in psychology (r = .55). It was, however, not related to major field of study (eta = .19, omega = .0) or any of the other background variables.

The number of undergraduate level terms of statistics was modestly related to perceived math ability (r = .26), to number of undergraduate terms of research methods (r = .30), to total number of terms of math (r = .25), and to major or minor in math as an undergraduate (point-biserial r = .23). It was not related to graduate major (eta = .18, omega = .0) or minor (eta = .27, omega = .0).

Coursework in Research Methods

The number of graduate level terms of research methods was related to major field of study (eta = .68, omega = .65), though not to minor field (eta = .32, omega = .0). In descending order, the major

fields ranked as follows: ECO, EXP, S-P, I-O, CL, and DEV. In general, CL subjects had fewer courses than NCL subjects (point-biserial r = -.27). Number of graduate level terms of research methods was related to low preference for clinical work as a career interest (r = .28), to number of statistics courses taken outside the Department of Psychology (r = .49), and to total number of graduate level statistics courses (r = .43).

The number of undergraduate level terms of research methods was not related to major field of study (eta = .24, omega = .0), but CL subjects tended to have fewer courses than NCL subjects (point-biserial r = -.23). Terms of undergraduate research methods was related to minor field of study (eta = .48, omega = .37). In descending order the minor fields ranked as follows: experimental psychology, quantitative and methodological psychology, other social sciences, physiological psychology/neuropsychology/biology, industrial-organizational psychology/business/management, social-personality psychology, none, ecological psychology, clinical psychology and developmental psychology (both tied). Undergraduate research methods was also related to high preference for teaching as a career interest (r = .23), to terms of undergraduate statistics (r = .30), to terms of math (r = .26), to terms of physics (r = .33), and to total terms of math and physical sciences (r = .27).

Coursework in Math and Physical Sciences

Total number of terms of math was related to perceived math ability (r = .39), to terms of undergraduate statistics (r = .25), to terms of undergraduate research methods (r = .26), to terms of

engineering (r = .33), to terms of physics (r = .45), and to terms of chemistry (r = .32). As to be expected, subjects who majored or minored in math had a greater number of math courses (point-biserial r = .57). Major field of study was not related to terms of math (eta = .15, omega = .0).

Total number of terms of engineering was related to terms of math (r = .33), to terms of physics (r = .57), and to undergraduate major or minor in math (point-biserial r = .35), but not to graduate major (eta = .16, omega = .0).

Total number of terms of physics was related to low preference for consulting as a career interest (r = .30), to perceived math ability (r = .34), to terms of undergraduate research methods (r = .23), to terms of math (r = .45), to terms of engineering (r = .57), to terms of chemistry (r = .62), to undergraduate major or minor in math (point-biserial r = .36), and to graduate minor (eta = .47, omega = .35). It was not, however, related to graduate major field of study (eta = .31, omega = .0). Also, men averaged more engineering and physics courses than women (point-biserial r's = .22, .24).

Number of terms of chemistry was only slightly if at all related to major field (eta = .36, omega = .26), though it was somewhat related to minor (eta = .49, omega = .38). In addition to the aforementioned relationships with terms of math (r = .32) and terms of physics (r = .62), terms of chemistry was related to several career interest preferences. It was related to high preference for teaching (r = .26), low preference for consulting (r = .26), and low preference for administrative work (r = .23).

Total terms of math and physical sciences was the sum of the number of terms of math, engineering, chemistry and physics. For this reason it was highly correlated with all these component variables. Outside of these, total terms of math and physical sciences was related to low preference for consulting ($r = .26$), to perceived math ability ($r = .36$), to terms of undergraduate research methods ($r = .27$), and to undergraduate major or minor in math (point biserial $r = .49$). It was not related to graduate major (eta = .23, omega = .0), and unrelated, or slightly related, to graduate minor (eta = .44, omega = .29).

Coursework in Philosophy

Coursework in philosophy was measured by four variables: total number of terms of philosophy, and three dichotomous variables indicating whether subjects had taken a course in logic, philosophy of science, and epistemology. None of these four variables showed any clear association with major or minor field of study, though the relationship between epistemology and major was the closest. Thus, for major field the degree of association with philosophy courses was respectively: total philosophy courses, eta = .32, omega = .04; for logic as the dependent variable, eta = .20; for philosophy of science as the dependent variable, eta = .28; for epistemology as the dependent variable, eta = .36. For minor field the corresponding measures of association are: eta = .37, omega = .15; eta = .26; eta = .27; eta = .23.

There were a few scattered relationships between the specific philosophy course variables and the other background and coursework variables. Having taken logic was related to age (point-biserial r =

.23) and number of terms of physics (point-biserial r = .22). Having taken philosophy of science was related to number of years of graduate work in psychology (point-biserial r = .22), and having taken epistemology was related to low preference for consulting (point-biserial r = .33), to undergraduate terms of research methods (point-biserial r = .26), and to total terms of math and physical sciences (point-biserial r = .23).

Question 2: What are their attitudes on statistical significance tests, psychology, and psychologists?

Cluster Analysis of the Attitude Questions

The ten attitude questions were used to assess individual differences in such qualities as naivete and optimism versus cynicism and skepticism concerning research and researchers in psychology, psychology as a science, and the use of significance tests. In addition to scores on these 10 items (numbered 33-42 in the questionnaire, but which are hereinafter designated as Attitude Questions 1 through 10), three composite attitude scales were formed on the basis of a cluster analysis. The three scales are hereinafter designated as Attitude Scales 1 through 3. The cluster analysis was based on the intercorrelation matrix (phi coefficients) of the attitude questions.

Initially clusters were formed by grouping together questions which had moderate intercorrelations. There were no high intercorrelations among the attitude questions, suggesting the possibility that individual questions had low reliabilities. In fact, this was one of the reasons for the cluster analysis and development of attitude scales (e.g., to increase reliability).

The three clusters or scales that were created were evaluated to see if they each were unidimensional. The three criteria for evaluation of unidimensionality are described by Hunter and Gerbing (1979). A unidimensional cluster has items that are: (1) internally consistent (moderate to high intercorrelations following a gradient consistent with their individual reliabilities); (2) externally consistent to within the considerations of sampling error (have a parallel

pattern of correlations with external variables or scales, which is proportional to their reliabilities); and (3) homogeneous in content (highly similar in meaning).

In order to evaluate the three attitudinal scales a multiple groups analysis was performed using the set of computer programs known as PACKAGE (Hunter and Gerbing, 1979; Hunter and Cohen, 1969). In addition to generating the intercorrelation matrix of items and scales, this analysis calculates the communalities for each item and the coefficient alphas for each scale. It corrects correlations between scales and their constituent items for inflation caused by a spurious common component, and it corrects the correlations between the scales and their external items for the attenuation due to measurement error.

Table 5 presents this corrected intercorrelation matrix. It also presents communalities for the questions used in the scales, and coefficient alphas for the scales. An item's communality is an estimate of its reliability given that it truly belongs to its assigned cluster, and coefficient alpha is an estimate of the reliability of the scale score, provided that a scale is unidimensional. Coefficient alpha will underestimate the reliability of a scale that is multidimensional and the correlations which are then corrected for attenuation will be suspect (Hunter and Gerbing, 1979).

The Three Attitude Scales

Attitude Scale 1, which includes Questions 1, 3 and 6, shows a smooth strong to weak gradient in correlations that is consistent with the gradient of communalities for these items. This is good evidence of internal consistency. An examination of the correlations with the

Table 5

THE ATTITUDE QUESTIONS AND SCALES:
INTERCORRELATIONS, COMMUNALITIES
AND COEFFICIENT ALPHAS*

	SCALE 1			SCALE 2			SCALE 3						
	1	3	6	2	4	5	8	10	7	9	SC1	SC2	SC3
1	36	27	19	-15	03	00	-02	08	13	-02	61	-09	04
3	27	19	11	03	08	05	03	-12	-06	-15	43	12	-07
6	19	11	10	-10	-03	-14	01	01	13	04	30	-20	01
2	-15	03	-10	30	25	20	09	15	-19	10	-17	55	18
4	03	08	-03	25	21	15	03	11	13	05	06	44	11
5	00	05	-14	20	15	14	-06	-04	08	03	-07	36	-07
8	-02	03	01	09	03	-06	45	42	-14	12	02	05	66
10	08	-12	01	15	11	-04	42	45	-05	09	-03	16	66
7	13	-06	13	-19	13	08	-14	-05	100	-11	15	02	-14
9	-02	-15	04	10	05	03	12	09	-11	100	-10	13	16
SC1	61	43	30	-17	06	-07	02	-03	15	-10	100	-13	-01
SC2	-09	12	-20	55	44	36	05	16	02	13	-13	100	16
SC3	04	-07	01	18	11	-07	66	66	-14	16	-01	16	100
Coefficient Alpha											.41	.43	.59

*Decimal points omitted from all correlations; communalities appear in the diagonal for the first eight items; correlations have been corrected for attenuation or spurious inflation.

external items and scales shows good evidence of parallelism, within the considerations of sampling error.[2] The content of the three questions appears highly similar, each concerned with the belief that significance tests prevent personal bias from entering into research conclusions or provide an objective means of evaluating hypotheses. For purposes of identification this scale is designated as: Significance Tests Equal Objectivity.

Attitude Scale 2, which includes Questions 2, 4 and 5, also shows a smooth strong-weak gradient of intercorrelations which is consistent with item communalities. This again is good evidence of internal consistency. Also, external consistency seems good. The content for the three items seems highly similar, each reflecting the opinion that psychologists as researchers tend to conclude what they started out with, either due to lack of caution in interpretation of data, or to biased use of significance tests. For purposes of identification this scale is designated as: Psychologists Conclude What They Want To.

Attitude Scale 3 includes Questions 8 and 10. Because it contains only two items it cannot be evaluated for a strong-weak gradient consistent with item communalities. However, the "product rule for internal consistency" can be applied (Hunter and Gerbing, 1979). The product rule in effect states that for any two items in a scale,

---

[2]For N = 80, the standard error of the difference of two independent correlations ranges from approximately .16 for population correlations of zero, to .15 for population correlations of .25, to .14 for population correlations of .50. Correction for attenuation on correlations with scale scores inflates the sampling error for these three scales from 23 percent to 36 percent, depending on the reliability of the scale.

the correlation between them should equal (to within sampling error considerations) the product of each of their corrected correlations with the scale score (e.g., their correlation with the true scale score). This is the case for Questions 8 and 10. External consistency seems very good for this scale, and the item content is highly similar. Both items reflect the belief that psychologists' research is on the shoddy side. Thus, this scale is named: Psychologists Do Shoddy Research.

Coefficient alphas for the three attitude scales are .41, .43, and .59 respectively, suggesting that the scales are not large enough in size (number of items) to raise their reliabilities to acceptable levels. Future work using attitude questions such as these should concentrate on measuring these dimensions with more items, perhaps even items that directly state the underlying attitudinal dimensions.

Two attitude questions, number 7 and 9, did not cluster with any of the attitude scales. Attitude Question 7, which assesses subjects' knowledge about the use of significance tests in other sciences, and Attitude Question 9, which measures the belief that psychology is not as rigorous a science as physics or biology, were left as individual items to be further analyzed.

It should be noted that none of the correlations among the three attitude scales and two extra attitude questions were statistically significant, though the direction of the correlations appeared to be consistent with the assigned meaning of the scales.

Individual Differences on the Attitude Scales

The response frequencies, means and standard deviations for the 10 attitude questions and three attitude scales are presented in Table 6. A look at the responses to the questions in Scale 1 shows that 50 percent of the sample disagreed with all three items, 34 percent disagreed with two of the three items, 13 percent disagreed with one of the items, and 4 percent disagreed with none, or agreed with them all. This suggests that the notion that significance tests equal objectivity is not widely or strongly believed among the graduate students in the sample.

Table 6

THE ATTITUDE QUESTIONS AND SCALES:
RESPONSE FREQUENCIES, MEANS, AND
STANDARD DEVIATIONS*

	1	3	6	2	4	5	8	10	7	9	SC1	SC2	SC3
Agree	29	8	19	57	40	44	67	54	37	47			
Disagree	51	72	61	23	40	36	13	26	43	32			
Score**													
0											40	11	10
1											27	18	19
2											10	30	51
3											3	21	--
Mean	.36	.10	.24	.71	.50	.55	.84	.68	.46	.60	.70	1.76	1.57
S.D.	.48	.30	.42	.45	.50	.50	.37	.47	.50	.49	.83	.98	.71

*Items coded: 0 = Disagree; 1 = Agree

**Frequency of subjects who agreed with this exact number of items.

Response frequencies for Scale 2 show that 26 percent of the sample agreed with all three items, 38 percent agreed with two items,

23 percent agreed with one item, and 14 percent agreed with none. Thus, it can be said that there is a fair amount of diversity in the extent to which subjects agree or disagree with the sentiment that psychologists conclude what they want to. At least half of the sample shows at least moderately strong agreement with this sentiment.

Response frequencies for Scale 3 show that 64% agreed with both items, 24 percent with one and 13 percent with none. Thus there seems to be a sizable number of subjects who agree that psychologists do shoddy work.

Question 3: How are these attitudes related to background and course-work characteristics?

Attitudes and Background Characteristics

Table 7 presents correlations or alternate measures of association for the three attitude scales and two unclustered attitude questions, and the background characteristics. Background variables include a dummy-coded CL versus NCL dichotomous variable and a dummy-coded dichotomous variable indicating whether the subject majored or minored in math as an undergraduate.

Only a scattered few relationships between these two sets of variables attain statistical significance or demonstrate a modest degree of association.

Females on the average were somewhat more likely to agree that statistical significance equals objectivity (Scale 1), and agree that most sciences make frequent use of significance tests (Question 7). Subjects who had a low preference for administrative work as a career goal were also more likely to agree with the items on Scale 1.

Subjects who agreed that psychologists conclude what they want to (Scale 2) were, on the average, older, with more years of graduate study in psychology. They also showed a higher preference for consulting and a lower preference for teaching as career goals.

Attitudes and Coursework Characteristics

Table 8 presents the correlations between the attitude variables and coursework characteristics. These correlations are for the most part small and fail to reach statistical significance. The largest of

Table 7

THE ATTITUDE VARIABLES AND BACKGROUND CHARACTERISTICS:
CORRELATIONS AND ALTERNATE MEASURES
OF ASSOCIATION*

Background Variables	Type of Measure	SC1	SC2	SC3	#7	#9
Major	eta	22	27	28	30	13
Major	omega	00	11	13	--	--
CL vs ncl**	r	08	-02	-07	-15	-11
Minor	eta	27	38	26	33	29
Minor	omega	00	19	00	--	--
UNDGR MATH MAJ/MIN**	r	-03	-08	05	03	09
Age	r	-06	29[b]	-12	-09	09
FEMALE vs male**	r	31[b]	17	11	30[b]	19
Yrs grad stdy-psych	r	-07	31[b]	16	-05	-08
Yrs grad stdy-other	r	-05	13	13	04	19
Research preference	r	01	-19	08	-08	17
Teaching preference	r	14	-26[a]	-06	-19	-08
Clinical preference	r	06	02	01	16	05
Consultg preference	r	-05	33[b]	-05	16	-21
Admin preference	r	-26[a]	19	-03	-11	-05
Math ability	r	03	-15	-01	-21	03

*Not corrected for attenuation; decimal points omitted

**Dummy-coded with upper case letters designating higher numerical coding; #'s 7, 9: 0 = Disagree, 1 = Agree

[a]Corresponding significance test: $p < .05$

[b]Corresponding significance test: $p < .01$

Table 8

THE ATTITUDE VARIABLES AND COURSEWORK CHARACTERISTICS: INTERCORRELATIONS*

Coursework Characteristics	ATTITUDE VARIABLES				
	SC1	SC2	SC3	#7	#9
Trms stat-psych	-09	10	-11[b]	-14	-02
Trms stat-other	-04	20[a]	30[b]	03	07
Trms stat-total	-05	26[a]	22[a]	-10	07
Yrs since stat	01	15	-21	02	-09
Trms undgr stat	06	-10[a]	01	01	-08
Res meth-grad	-01	27[a]	12	-04	11
Res meth-undgr	-08	-06	06	11	18
Trms math	09	04	-03	-16	11
Trms engineer	-11	09	13	-18[a]	-01
Trms physics	-02[a]	-04	00	-23[a]	06
Trms chem	22[a]	-03	-09	-13	16
Trms sci-total	09	06	02	-17	12
Trms philo	-12	09	04	-12[a]	-06
LOGIC**	-18	09	04	-27[a]	-02
PHILO OF SCI**	-20	12	05	-15	-01
EPIST**	-14	-01	15	10	-02

*Not corrected for attenuation; decimal points omitted

**Dummy coded with upper case letters designating higher numerical coding; #'s 7,9: 0 = Disagree, 1 = Agree

[a] $p < .05$ [b] $p < .01$

these correlations indicate that subjects who agreed that psychologists conclude what they want to (Scale 2) and that psychologists do shoddy work (Scale 3) took more total terms of graduate level statistics and more terms of statistics outside the Department of Psychology. Subjects who agreed that most sciences make frequent use of significance tests (#7) took fewer terms of physics and were less likely to have taken a philosophy course in logic. Also, subjects who agreed that psychologists conclude what they want to (Scale 2) took more graduate terms

of research methods and subjects who agree that significance tests equal objectivity (Scale 1) may have taken more terms of chemistry.

In summary it can be said that there were no clearcut, systematic patterns of relationships between the attitude variables and the other biographical characteristics.

Question 4: What kind of research conclusions do subjects believe are valid in Problems 1 and 2? What rules or strategies can be inferred from these responses?

The Profile Analyses

In order to describe the patterns of responses and classify the different response strategies used by subjects, two separate profile or pattern analyses were conducted (Nunnally, 1967). The method used for grouping the response patterns in each problem began with an exploratory Q-type factor analysis and blind clustering procedure (Stephenson, 1953; Kim, 1975). This factor analysis--a principle components analysis with communalities, followed by a series of varimax rotations--was performed on the matrix of Q-type correlations between subjects.

The usual correlation matrix (R-type) submitted for the usual factor analysis (R-type) is the matrix of intercorrelations among variables. However, since the purpose of the present analysis was to group or classify subjects according to similarity in patterns of responses across 36 variables, the matrix submitted for analysis was the matrix of intercorrelations between subjects. Correlations between persons are thought to be a good measure of profile similarity as long as profile level and profile scatter or dispersion are not critical considerations (Nunnally, 1967; Neufeld, 1977). If profiles are to be grouped according to the shape of the profile, that is, "with respect to the interrelationships among the measures, and not to similarity in the measures' absolute values or intermeasure variance," then a correlation coefficient is a good measure of similarity (Neufeld, 1977, p. 154).

Neither absolute values not intermeasure variance were considered of critical importance given the dichotomous nature of the variables.

In order to obtain the two Q-type correlation matrices, the two data matrices submitted to a preliminary correlation computer program had to be transposed or inverted. Subjects, or cases, instead of variables had to form the columns in the data matrix of each problem, while variables, usually arrayed by columns, had to be arrayed by rows.

After the varimax rotations were performed, groups of subjects were clustered into profile groups (PG's) using a blind clustering procedure. One PG or cluster of subjects was formed for each factor created in the varimax procedure. The subjects forming a particular group were those who had their highest loading on that factor. Thus, two sets of groups were created, one for each problem.

One limitation of this technique is that the blind clustering procedure assumes that variables in the factor analysis (in this case subjects) are dimensional in nature and could thus have been reversed scored, or reflected. Subjects as variables are not dimensional in this way and it is not meaningful to reflect them. Thus, for Problem 1 three subjects were placed into groups based on their highest factor loading being negative and for Problem 2 there were five such subjects. These subjects had to be reclassified.

The exploratory Q-type factor analyses and blind clustering procedure was used as a method of initially classifying a large number of subjects into a set of PG's. This provisional set of groups was then inspected for homogeneity. Specifically, the data for each of the eleven groups in Problem 1 and the eleven groups in Problem 2 was assembled separately and visually inspected for similarity. A twelfth

group was immediately added to each set of PG's. It was formed from those subjects who had to be excluded from the factor analyses because they showed no variation across items (e.g., they judged all items as invalid), and hence could not be correlated.

After visual inspection of the data modifications were made on PG's that did not have highly similar response patterns for all subjects. Several groups had to be subdivided into smaller subgroups, and a few subjects were removed from groups because their response patterns were only partially similar to those of the other members of the group. Attempts were made to find alternate groups for these subjects. If this proved impossible, as it did for several subjects in both problems, then they were placed in a residual group of unclassifiable subjects.

With minor modifications the PG's formed in the factor analyses and blind clustering procedures proved satisfactory. Sixteen distinct PG's were formed for Problem 1, and fifteen were formed for Problem 2 (this includes the residual group for each problem). Each of these appeared to represent a distinct response pattern and decision rule for that problem. The profile groups for Problems 1 and 2 are presented in Appendix B.

The Typologies and Miscellaneous Variables

The large number of PG's required to adequately classify the different response patterns in Problems 1 and 2 made it advisable to search for a system to collapse these into a smaller number of more general groupings or types. This would allow comparison of the two

profile analyses at a more abstract and general level, which was important because the fine-grained comparison of the two profile analyses using PG membership showed marked inconsistencies across problems. A smaller number of groups would also serve to facilitate comparisons between the profile analyses and individual difference variables. What was needed was some general, more abstract classification scheme which would function to combine the PG's into a much smaller number of profile types. An attempt to do this by grouping together PG's with similar looking patterns proved unsatisfactory for both problems. This was because the PG's were distinct enough such that any combination of several of them proved so diverse and amorphous that as a type it was unidentifiable.

The Typologies

The procedure to create general types from the PG sets that was ultimately developed was based on three dichotomous or Bernoulli variables, each indicating whether or not a particular factor (or factors) was used by subjects in a PG.

The first Bernoulli variable indicated whether or not a particular PG used either the Population Generality Factor, the Task Generality Factor, or both, in their response rule. These two factors were combined into one Bernoulli variable because subjects tended to use them together. The second Bernoulli variable indicated whether or not the Format Factor was a consideration in a PG's response. The third Bernoulli indicated whether or not a PG used the Theoretical Generality Factor in their response rule. Thus each PG was rated either no or yes (scored - or +; or dummy-coded 0, 1 for computer

analyses) for each of the three Bernoulli indicators depending on whether or not that factor(s) had been a consideration in the responses of the PG.

The Bernoulli indicators were used to set up an eight category typology representing all possible combinations of the three variables. Thus, in addition to receiving a score on the three Bernoulli variables, each PG was assigned a typological classification from 1 to 8, as shown in Tables 9 and 10. (Subjects in the residual PG's were given the individual assignments.)

Table 9

COMPOSITION OF TYPOLOGY 1 (Problem 1)

Type	Bernoulli's* PT1	TE1	OT1	PG's Included	N
1	+	+	+	6, 9, 11, 12, 13	22
2	+	+	-	7, 8, 14	17
3	+	-	+	10	2
4	-	+	+	1, 2, 3	22
5	+	-	-	5	1
6	-	+	-	4	11
7	-	-	+	None	0
8	-	-	-	15	5
N (-)	38	8	34		
N (+)	42	72	46		

*PT1 = Population or Task Factors
TE1 = Format (Tentativity) Factor
OT1 = Theoretical Factor

Table 10

COMPOSITION OF TYPOLOGY 2 (Problem 2)

Type	Bernoulli's* PT2	TE2	OT2	PG's Included**	N
1	+	+	+	2, 6, 7, 10, 11	20
2	+	+	-	3, 4, 5, 9, 10, 11	28
3	+	-	+	2, 7, 8, 12	6
4	-	+	+	2	1
5	+	-	-	None	0
6	-	+	-	1, 4	19
7	-	-	+	None	0
8	-	-	-	13	6

*PT2 = Population or Task Factors

TE2 = Format (Tentativity) Factor

OT2 = Theoretical Factor

**Some PG's appear more than once if they were subdivided when assigned to types.

The interrelationships of the three Bernoulli indicators were examined for each problem by setting up contingency tables and computing Chi-squares and phi coefficients. For Problem 1, phi coefficients ranged from -.01 (PT1 X OT1), to .10 (PT1 X TE1), to .22 (TE1 X OT1), none of which attained conventional significance levels. For Problem 2, phi coefficients ranged from -.16 (TE2 X OT2), to .17 (PT 2 X TE2), to .47 (PT2 X OT2). The PT2 X OT2 level of association was clearly significant ($p < .0001$). The contingency table for these two Bernoulli's clearly showed that on Problem 2 every subject that used the Theoretical Generality Factor also used the Population or Task Factors, and that subjects that used the Population or Task Factors were more likely to use the Theoretical Generality Factor than those that didn't.

The Miscellaneous Variables

As part of each problem all subjects were asked to indicate which of the conclusions was the best, and if they were familiar with the area of research involved in the problem. Also, a calculation of the number of minutes it took the subject to complete each problem was made, based on entries as to the starting and stopping time for that problem.

The frequency distribution for the best choice for Problem 1 showed that #3 and #9 were the two most frequent choices. Item #3 was the most popular, with 44 percent of the sample choosing it. The conservative choices 37, 1, 2, and 3 accounted for 61 percent of the sample, and choices 4 through 9 accounted for another 23 percent.

The frequency distribution for the best choice for Problem 2 showed that #3 was the most popular item, with 39 percent of the sample choosing it. Second and third in frequency were #3 and #9. Again, the conservative choices 37, 1, 2, and 3 accounted for 61 percent of the sample.

Only three subjects in the sample indicated they were familiar with the research in Problem 1. However 29 subjects (36 percent) indicated they were familiar with the research in Problem 2.

The mean number of minutes taken to complete Problem 1 was 10.5, with a standard deviation of 3.5 and a range of 15. The distribution was positively skewed (.41). The mean number of minutes taken to complete Problem 2 was 8.0, with a standard deviation of 3.0 and a range of 16. The distribution was also positively skewed (.49).

In order to determine if prior familiarity with the problem

content area or speed in completing the problem were related at all to the PG's, the types, or the Bernoulli indicators for that problem, a series of cross-tabulations, one-way analyses of variance, and point-biserial correlations were conducted. Using measures of association (eta, omega, phi, and r) in addition to significance tests, it was reasonably clear that all of these relationships were trivial in magnitude.

Question 5: What kinds of generalizations do subjects make in Problems 1 and 2? Are these related to the miscellaneous variables for each problem?

The Generalization Scales

In order to determine the degree to which subjects generalized on the four dimensions (factors) described above, a series of five summary or Generalization scales were developed. For each problem subjects' scores for the five scales were computed and made available for analysis.

Scale 1 (TenQ) measured a subject's preference for conclusions with a tentative qualified format over a tentative (unqualified) format. This was accomplished by subtracting the number of tentative (unqualified) conclusions judged valid from the number of tentative qualified conclusions judged valid. Positive scores reflected a preference for the qualified items over the unqualified items and negative scores indicated the opposite preference. Preference for the qualified over the unqualified was a measure of the misinterpretation described above as "confusing a conditional probability with an absolute probability" which is manifested in the belief that the p value identifies the probability that the results of an experiment are due to chance.

Scale 2 (Theo) measured preference for theoretical over operational conclusions. This was accomplished by subtracting the number of operational items accepted from the number of theoretical items accepted. Higher scores represented a greater preference for theoretical over operational items.

Scale 3 (Pop) measured the preference for population general and

college sophomore items over population specific items. This was accomplished by subtracting the number of population specific conclusions judged valid from the number of population general and college sophomore items judged valid. For this scale the higher the score the greater the preference for population general and college sophomore items.

Scale 4 (Task) measured the preference for task unqualified items over task specific items. This was accomplished by subtracting the latter from the former. For Scale 4, the higher the score the greater the preference for accepting task unqualified over task specific items. Scale 4 was formulated to help to gauge the degree to which subjects were willing to generalize on the basis of a certain experimental situation.

Scale 5 (Tent) measured the general preference for tentative (either level) over nontentative items. This was accomplished by subtracting the number of nontentative conclusions judged valid from the number of tentative conclusions judged valid. Unlike the other four scales, this scale was set up such that higher scores represented more cautious decision strategies.

Means, standard deviations, and intercorrelations for the five Generalization Scales are presented in Table 11. For both Problems 1 and 2, Generalization Scales 3 and 5 are highly correlated. Thus, the degree of preference for tentative (either level) over nontentative (Scale 5) is strongly associated with the degree of preference for population general and college sophomore over population specific (Scale 3). The only other correlation that was statistically significant

Table 11

GENERALIZATION SCALES: MEANS, STANDARD DEVIATIONS, AND INTERCORRELATIONS

	Problem 1						
	1	2	3	4	5	Mean	S.D.
1(TenQ)	1.00					.6	2.9
2(Theo)	.07	1.00				-3.8	4.0
3(Pop)	.04	-.16	1.00			1.4	3.8
4(Task)	.09	-.04	-.08	1.00		-2.0	2.5
5(Tent)	.03	.01	.65[b]	.07	1.00	8.8	6.3
	Problem 2						
1(TenQ)	1.00					.4	3.7
2(Theo)	.29[b]	1.00				-1.8	3.2
3(Pop)	.10	.14	1.00			1.0	3.7
4(Task)	-.08	-.07	.05	1.00		-2.9	3.3
5(Tent)	-.08	.10	.71[b]	.11	1.00	8.3	6.9

[a] $p < .05$ [b] $p < .01$

was that between Scale 1 and Scale 2 for Problem 2 only. Because this relationship was not replicated in Problem 1, it is probable that this correlation is the result of sampling error.

With the exception of the strong relationship between Scale 5 and Scale 3, it appears that the generalizations are independent from one another.

The Generalization Scales and Miscellaneous Variables

In order to determine if prior familiarity with the problem content area or speed in completing the problem were related at all to the five Generalization Scales for that problem, a series of point-biserial and Pearson correlations were computed. These are presented in Table 12.

Table 12

THE GENERALIZATION SCALES AND MISCELLANEOUS VARIABLES: INTERCORRELATIONS**

	Problem 1					Problem 2				
	1	2	3	4	5	1	2	3	4	5
Familiar*	-05	-11	02	29[b]	06	-02	-02	02	00	-07
Minutes	-22[a]	02	-05	-04	-08	-11	20	05	11	01

*Coded: 0 = Yes; 1 = No

**Decimal points are omitted

[a] $p < .05$ [b] $p < .01$

Only two of these correlations are statistically significant and both of these were found only in Problem 1. The parallel correlations for Problem 2 do not replicate these relationships, hence it is most likely that these two statistically significant correlations are the result of sampling error.

These two isolated relationships between two of the miscellaneous variables and the five Generalization Scales indicate that familiarity with the content area of the problem, or amount of time spent in completing a problem are probably not related to the generalizations. This is consistent with previous findings which indicated that these two miscellaneous variables were not all related to the PG's, the types, or the Bernoulli indicators.

Question 6: To what extent are subjects consistent in response strategies and generalizations in Problems 1 and 2?

Consistency in Response Strategies

In order to determine the extent to which response strategies were consistent across Problems 1 and 2, a joint frequency distribution of membership in the two sets of PG's was constructed (see Table 13).

Table 13

PROFILE GROUPS FOR PROBLEMS 1 AND 2:
JOINT FREQUENCY DISTRIBUTION

Prob 1 PG's	Problem 2 PG's															Row Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	1	2	0	0	0	1	0	0	0	0	1	1	0	0	0	6
2	0	3	1	3	0	4	0	0	1	0	0	0	0	0	0	12
3	1	0	0	0	0	0	0	0	0	0	0	0	0	2	1	4
4	5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	6
5	3	0	1	0	0	0	1	0	0	0	0	0	0	0	0	5
6	0	2	2	0	0	0	1	0	2	0	0	0	0	0	0	7
7	0	0	3	0	0	1	0	0	0	1	0	0	0	1	0	6
8	3	0	2	1	0	0	0	0	0	0	0	0	1	0	0	7
9	1	0	0	0	0	1	0	0	0	0	2	0	0	0	0	4
10	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	3
11	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
12	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	4
13	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2
14	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	2
15	1	0	1	0	0	0	0	0	0	0	0	0	4	0	0	6
16	1	0	1	0	0	0	1	0	0	1	1	0	0	0	0	5
Column Total	16	9	12	6	3	7	3	1	3	3	4	2	6	3	2	

This rather large contingency table showed that subjects in a PG from one problem were usually dispersed among several PG's in the other problem. For example, the 12 subjects in PG 2 from Problem 1 were members of five different PG's in Problem 2. If response

strategies were completely consistent across problems, then members of a particular PG in one problem would all be members of the same PG in the other problem. This was rarely the case. Thus, there appeared to be a substantial degree of inconsistency in response strategies across problems at this level of analysis.

In order to investigate the inconsistencies in PG membership for Problems 1 and 2 at a more general level of analysis, the intercorrelations of the Bernoulli indicator variables were examined. The phi coefficients for the three Bernoulli variables are presented in Table 14.

Table 14

THE BERNOULLI INDICATORS FOR PROBLEMS 1 AND 2: INTERCORRELATIONS*

	Problem 2		
Problem 1	PT2	TE2	OT2
PT1	.22[a]	.02	-.09
TE1	.13	.68[b]	-.04
OT1	.35[b]	.06	.38[b]

*Phi coefficients; PT = Population or Task Factors, TE = Format Factor, OT = Theoretical Factor.

[a] $p < .05$ [b] $p < .01$

The greatest consistency can be seen for the TE indicator variable, suggesting that whether or not a subject used the Format (Tentativity) Factor was moderately consistent from problem to problem. Only a modest degree of consistency seemed to be the case for the Theoretical Generality Factor. Of the 28 subjects who were

inconsistent in their use of this factor, 24 used it as factor in Problem 1 but not in Problem 2.

Finally, there appeared to be a small amount of consistency in whether or not the Population or Task Factors were used. Of the 29 subjects who were inconsistent in their use of these factors, 20 subjects who had not used either in Problem 1 used one or both in Problem 2.

From the analysis of PG membership and the intercorrelations of Bernoulli indicators it appears that only a moderate degree of consistency exists in response strategies in Problems 1 and 2.

Consistency in Generalizations

In order to determine the degree of consistency in generalizations across Problems 1 and 2, the Generalization Scale scores were correlated. These are presented in Table 15.

Table 15

THE GENERALIZATION SCALES:
CORRELATIONS ACROSS PROBLEMS

	Problem 2				
Problem 1	1	2	3	4	5
1 (TenQ)	.60[b]	.20	.04	-.18	-.06
2 (Theo)	.15	.26[a]	.00	.00	.00
3 (Pop)	-.02	-.09	.48[b]	.14	.37[b]
4 (Task)	.08	.04	.06	.40[b]	.03
5 (Tent)	-.05	-.03	.50[b]	.17	.71[b]

[a] $p < .05$ [b] $p < .01$

An examination of the correlations in the diagonal of Table 15

shows that scores on Scale 5 and Scale 1 had the highest consistency. Scale 5 measures the degree of preference for tentative (either level) over nontentative items. Scale 1 measures preference for tentative qualified items over tentative (unqualified) items. Thus, the greatest consistency across problems appears to be in the use of the Format Factor. This finding is consistent with the analysis based on inter-correlations among Bernoulli variables.

A moderate degree of consistency was also evident for Scales 3 and 4. Scale 3 measures the degree of preference for population general and college sophomore items over population specific items. Scale 4 measures the preference for task unqualified items over task specific items. Thus, the two Generalization Scales that involve the Population and Task Generality Factors show a moderate degree of consistency. This is somewhat more than what was found when Bernoulli indicator variables were correlated.

Only a small degree of consistency was evident for Scale 2, a measure of preference for theoretical over operational items. This degree of consistency was less than what was evident in the analysis of the corresponding Bernoulli indicator variables.

Summary: Consistency in Problems 1 and 2

Consistency in response strategies and in generalizations varies depending on the type of factor involved. Subjects on the whole showed greatest consistency in their use of the Format Factor. Less consistency was evident for the other three factors. It should also be noted that an examination of each subjects' actual responses to the two problems shows that some subjects seemed quite consistent

while other appeared quite inconsistent. To a large extent this inconsistency is probably the result of a random error process. A small part of it, however, may have been due to real or imagined logical differences in problems.

Question 7: What strategies do subjects use to make judgments about strength of relationship and personal assurance, for the two statistics presented in Problems 3 through 6?

Profile Analysis - Problems 3 through 6

In order to ascertain the response strategies used by subjects in making judgments of personal assurance that the null hypothesis is false and strength of relationship, the responses given to Problems 3 through 6 were subject to a profile analysis. The classification method for this was identical to that used in Problems 1 and 2. It was derived from an exploratory Q-type factor analysis of the intercorrelation matrix of subjects.

First Attempt: The Directional Preference Scores

The first attempt at generating a set of profile groups was done using the 24 weighted directional preference scores for these four problems. A first run produced a set of PG's that was only partially satisfactory. Inspection of the data for about half the PG's showed the absence of any clear and definitive commonality across subjects, possibly because there was a lot of randomness and error in individual responses.

Two strategies were used to attempt a satisfactory completion of the profile analysis of these 24 scores per subject. First, a second exploratory factor analysis was done using those subjects that were residual from the first run (e.g., that could not satisfactorily be included in any acceptable PG). This second factor analysis was only somewhat helpful. It was only able to generate two additional PG's that

on inspection of the data were at least marginally acceptable. Between these two factor analyses and blind grouping runs about 75 percent of the original 80 subjects were able to be placed in acceptable PG's. Thus, a significant percentage of subjects remained in groups that were unsatisfactory or only marginally acceptable.

It should be remembered that approximately half the sample had failed to follow instructions for this portion of the questionnaire, calling into question the reliability of these responses. Also, this was the section of the questionnaire that was the most technical in nature, hence the most difficult for subjects. For these reasons it was suspected that responses to these items contained a good deal of randomness. With this suspicion in mind a second strategy was adopted to attempt to complete a satisfactory profile analysis for directional preference scores. For those questionable or unsatisfactory PG's in the second factor analysis a series of mean vector scores (MVS's) was calculated for each PG. It was hoped that averaging across subjects would highlight, amidst the noise of individual subject's responses, a clear-cut response rule for that group. This, unfortunately, did not turn out to be the case.

Second Attempt: The Problem Choice Responses

Because the profile analysis of the weighted directional preference scores was not completely satisfactory, it was abandoned and a profile analysis of the eight choice responses was performed.[3] The use of these

---

[3]There were eight comparisons in Problems 3 through 6--each of the four problems had two subsections. Each of the eight comparisons had a choice response and three directional preference scores, indicating why subjects had made the choice they did.

responses provided a more reliable data base for a profile analysis, though it meant that response strategies had to be inferred, e.g., the reason for making a choice had to be inferred from the overall pattern of choices, though this could be checked against the directional preference scores for that choice. The same exploratory factor analysis and blind clustering procedures were used and a completely new set of PG's were formed. With very few modifications these groups proved satisfactory.

For each of the PG's a choice response rule was inferred. In order to check the validity of these inferences the weighted directional preference scores were assembled for the subjects in each PG and 24 MVS's were calculated per group. The MVS's were calculated to serve as a summary against which the inferred choice response rules could be compared, and to help clarify cases in which there was some ambiguity in these rules. The description of the profile groups for Problems 3 through 6 are presented in Appendix C.

Typology 3 - Problems 3 through 6

The large number of PG's required to adequately classify the different response patterns in Problems 3 through 6 made it advisable to develop a more general, higher-order classification scheme with a smaller number of categories.

The typology developed was based on two Bernoulli variables. The first of these variables (J) indicated whether a particular PG had differentiated between the two different judgment types--personal assurance and strength of relationship--that is, whether they used different decision strategies for the two types of judgment. The

second variable (S) indicated whether a particular PG had differentiated between the two statistics, t and r. Thus, each PG was rated no or yes (- or +) for each of the Bernoulli variables, depending on whether their response rules differed for the two components of that Bernoulli.

A four category typology representing all possible combinations of the two variables was formed. This is presented in Table 16.

Table 16

COMPOSITION OF TYPOLOGY 3
(PROBLEMS 3 THROUGH 6)*

Type	Bernoulli's J	Bernoulli's S	PG's Included	N
1	+	+	1, 8, 9, 10, 11, 12, 14, 15	31
2	-	+	13	3
3	+	-	5, 6, 7	8
4	-	-	2, 3, 4, 16	35
N (-)	38	43		
N (+)	39	34		

*PG 17 was not included

A crosstabulation between the two Bernoulli indicators showed that they were highly related (phi = .75). Thus, PG's that differentiated on one Bernoulli were likely to also differentiate on the other."

The Recode Variable

In order to determine if there was any relationship between membership in the PG's and types of Problems 3 through 6, and whether a subject's directional preference scores had been recoded (e.g., whether the instructions for parts of Problems 3 through 6 had been misunderstood),

a series of crosstabulations were performed. These were done between the dummy-coded Recode variable (see p. 30), and membership in the PG's, the types, and scores on the two Bernoulli variables. None of the Chi-squares that were computed approached the .05 significance level, though the one between the Recode variable and PG membership was the closest ($p = .20$). Eta, with Recode as the dependent variable, was .49 for PG membership, and .14 for Typology 3 membership. Phi coefficients between the Recode variable and the two Bernoulli's were both less than .07.

Question 8: In what ways do the rules for judging personal assurance and strength of relationship differ from ideal or correct rules?

The Error Variables

In order to analyze the extent to which subjects used incorrect rules in making judgments of personal assurance and strength of relationship, a series of five Error Variables (EV's) was established. Each of the five EV's was a dichotomous or Bernoulli variable which indicated whether or not one of five different errors had been committed.

EV 1 indicated whether or not a particular PG committed "the large N fallacy;" that is, whether they used large N in making judgments of personal assurance. As noted above, N is irrelevant in making judgments about the falsehood of the null hypothesis when p is provided. PG's that used large N as a decision rule in one or both personal assurance problems (3 and 5) were scored positive on this indicator. EV 1 was the only indicator in which positive or high scores denoted error. In all others positive scores denoted use of the correct rule.

EV 2 indicated whether or not any particular PG had correctly used low p as the only or primary factor in making personal assurance judgments when t was the statistic presented. EV 3 indicated whether or not any particular PG had correctly used low p as the only or primary factor in making personal assurance judgments when r was the statistic presented.

EV 4 indicated whether or not any particular PG had used r as the only or primary factor in judging strength of relationship when r was the statistic presented.

EV 5 indicated whether or not any particular PG had used small N as a primary factor in judging strength of relationship when t was the statistic presented. Thus, both EV 1 and EV 5 measured (though scored in opposite directions) the "confusion about the relationship between N and p" and the "confusion about the relationship between N, p and strength of association" described above.

Frequencies of responses, intercorrelations and PG's making errors for the EV's are presented in Table 17.

Table 17

ERROR VARIABLES: FREQUENCIES, INTERCORRELATIONS, AND PG'S MAKING ERRORS*

	Phi coefficients					Frequencies		PG's Making Error
	EV1	EV2	EV3	EV4	EV5	-(0)	+(1)	
EV 1	1.00					56	21	3,6,7,13-15
EV 2	-.75[b]	1.00				31	46	2,3,5-8,12-14
EV 3	-.23[a]	.48[b]	1.00			37	40	2-8,10,12,15
EV 4	-.41[b]	.31[b]	-.20	1.00		24	53	3,5,12-16
EV 5	-.18	-.30[b]	-.23[a]	.29[a]	1.00	65	12	1,3-6,8-10,12-16

*Based on 77 subjects in PG's 1 through 16

[a] $p < .05$ [b] $p < .05$

The most frequent error made by subjects was EV 5; 84 percent of the subjects failed to use small N as a primary factor in judging strength of relationship when t was the statistic presented. This was followed by EV 3 and EV 2, where 48 percent and 40 percent of the subjects, respectively, failed to use low p as the primary factor in personal assurance judgments with r and t, respectively. EV 4 shows that 31 percent of the subjects failed to use r as the primary factor in judging strength of relationship when r was presented. The large N

fallacy, as measured by EV 1, was the least frequent error; 27 percent of the subjects used large N in one or more judgements of personal assurance.

An examination of the interrelationships among the EV's shows the strongest association to be between EV 1 and EV 2. EV 1 and EV 2 show a strong inverse relationship (however, it should be remembered that EV 1 was scored in the opposite direction from EV's 2 through 5). The contingency table for these two variables showed that a subject that made an error on EV 1 was more likely to make an error on EV 5, and that a subject that made an error on EV 5 was more likely to be in error than to be correct on EV 1.

EV 2 and EV 3 showed a moderately strong degree of association. The contingency table showed that errors on one were associated with a greater likelihood of errors on the other.

EV 1 and EV 4 also showed a moderately strong (inverse, because of EV 1's reversed scoring) association. Here the contingency table showed a strong tendency for subjects who used a correct rule in one case to use a correct rule in the other. The tendency for the obverse (errors on one associated with errors on the other) was not as strong.

The relationship between EV 2 and EV's 4 and 5, and between EV 4 and EV 5, showed a modest, statistically significant degree of association. EV 2 had a direct relationship to EV 4 and an inverse relationship to EV 5, and EV 4 and EV 5 had a direct relationship. The contingency table for EV 2 and EV 4 showed a tendency for subjects who used the correct rule on one to do the same on the other. The obverse, however, does not show as clear a tendency. The contingency table for EV 2 and

EV 5 shows that subjects who used the correct rule on one were more likely to err on the other. The obverse was also evident from the contingency table.

The contingency table for EV 4 and EV 5 showed that all subjects who erred on EV 4, erred on EV 5, too. It also showed that subjects who erred on EV 5 were more likely to err than be correct on EV 4.

There were small, but statistically significant inverse relationships between EV's 1 and 3, and EV's 3 and 5. Subjects that erred on EV 1 were more likely to err on EV 3 (though not vice versa), and subjects that erred on EV 3 were more likely to err on EV 5 (though again not vice versa).

Thus, the strongest relationship was between committing the large N fallacy and failing to use low p in assurance judgments with t, possibly because in many cases large N was used instead of p. The other two outstanding relationships were between the two EV's measuring the use of low p as the primary factor in the two personal assurance judgments (t and r), and between the large N fallacy and the use of r in judging strength of relationship. Thus, most of the strongest relationships between the EV's were between those that overlapped in measuring specific problems, such that certain erroneous strategies would get picked up in both (e.g., EV 1 and EV's 2 and 3), or between EV's that measured the same judgment type (EV's 2 and 3; EV's 4 and 5) or the same statistic (EV's 2 and 4), where subjects who used the same erroneous strategy in both problems would be picked up in two EV's.

The Error Variables and Generalization Scales

In order to determine if any of the EV's were related to the Generalization Scales a series of correlations were calculated. These are presented in Table 18. It is clear from these correlations that all of the relationships were trivial.

Table 18

THE ERROR VARIABLES AND GENERALIZATION SCALES: INTERCORRELATIONS*

	Problem 1					Problem 2				
	1	2	3	4	5	1	2	3	4	5
EV 1	-10	-18	09	-03	13	-15	02	06	07	11
EV 2	11	16	-05	09	-05	11	03	03	-09	-03
EV 3	01	00	-05	04	-05	08	08	00	-14	-02
EV 4	-07	04	-18	09	-20	03	06	-06	04	-11
EV 5	-15	-07	-10	-04	-09	-11	-05	-13	10	-13

*Decimal points omitted in point-biserial correlations

$^{a}p < .05$ $^{b}p < .01$

Question 9: What biographical characteristics (background, course-work, attitudes) are related to generalizations and errors in Problems 1 through 6?

Background Variables and Generalization Scales

Correlations and alternate measures of association between the background variables and Generalization Scales are presented in Table 19. Only seven out of 160 relationships are significant at the .05 level or better. This is about what could be expected as a result of sampling error even if all relationships were zero. However, a number of relationships are replicated across problems, which is evidence that these relationships are not the result of sampling error. Also, a number of measures of association which bordered on but did not attain statistical significance were replicated across problems, suggesting that these too were nontrivial.

Of the seven statistically significant correlations only three show clear evidence of replication. Whether a subject was in the clinical program (CL vs. NCL) or had a high career preference for clinical work was modestly related to preference for tentative over nontentative items(Scale 5, a more cautious strategy). The third statistically significant relationship that replicated was the relationship between number of years of graduate study in psychology and Scale 1, which measured preference for tentative qualified over tentative (unqualified) items. Thus, more advanced students were more likely to believe the p value represents the probability of error due to chance.

A number of statistically significant correlations did not replicate across problems. Since it is unlikely that biographical

Table 19

THE BACKGROUND VARIABLES AND GENERALIZATION SCALES:
CORRELATIONS AND ALTERNATE MEASURES
OF ASSOCIATION**

Background Variables	Problem 1					Problem 2				
	1	2	3	4	5	1	2	3	4	5
Major (eta)	27	25	17	15	24	23	14	14	14	34
Major (omega)	09	00	00	00	00	00	00	00	00	10
CL vs. ncl*	02	-19	05	07	21	-11	07	07	-03	30[b]
Minor (eta)	41	30	37	44	42	41	39	33	37	42
Minor (omega)	25	00	17	29	27	24	21	00	16	27
Undgr major (eta)	19	14	31	46[a]	28	30	29	28	17	26
Undgr major (omega)	00	00	11	37[a]	00	06	00	00	00	00
Undgr minor (eta)	31	28	38	33	24	37	28	35	27	19
Undgr minor (omega)	00	00	18	11	00	21	00	16	00	00
UNDGR MATH MAJ/MIN*	06	-15	06	-12	-12	-11	11	-22[a]	-07	-12
Age	04	17	00	03	02	18	06	04	07	03
FEMALE vs. male*	09	10	-03	03	07	-03	10	-02	-04	00
Yrs grad stdy-psych	29[b]	19	07	03	01	21	-06	-05	00	-13
Yrs grad stdy-other	21	20	12	-03	12	13	08	-06	01	03
Research pref[1]	06	-15	05	03	18	-09	17	06	06	21
Teaching pref[1]	-02	-12	-07	-06	08	-09	-08	-07	05	09
Clinical pref[1]	11	26[a]	01	-07	-17	16	02	-04	-08	-29[b]
Consulting pref[1]	-15	03	-06	-04	-08	07	-07	08	-07	07
Admin pref[1]	-14	-13	03	18	06	-14	-09	01	07	10
Math ability	06	-26[a]	-20	06	-22	-12	01	-21	-06	-14

*Dummy-coded with upper case letters designating higher numerical coding

**Decimal points omitted; statistics presented are correlations except where indicated

[1]Low preference scored high

[a]$p < .05$ [b]$p < .01$

variables would be related to a Generalization Scale in one but not the other problem, these nonreplicated correlations were probably the result of sampling error.

There were five relationships between the background variables and Generalization Scales which while not statistically significant did replicate. First, low research career preference was associated with a greater preference for tentative over nontentative items (Scale 5). Second, subjects who perceived their math ability as better were less likely to prefer college sophomore or population general over population specific items (Scale 3).

The remaining three replicated relationships were between graduate minor and Scales 1, 4 and 5. An examination of the mean scores for each minor field category showed that no one minor or set of minors did consistently better or worse on these scales.

Coursework Characteristics and Generalization Scales

Correlations between the coursework characteristics and Generalization Scales are presented in Table 20. Of 160 correlations, 15 are significant at the .05 level or better. This is somewhat more than would be expected on the basis of sampling fluctuations alone.

Of these 15 correlations, seven replicated across problems. There were no correlations that were not statistically significant which replicated across both problems.

Of those that replicated, number of terms of statistics taken in the Department of Psychology and total terms of graduate level statistics were related to preference for tentative qualified over tentative unqualified (Scale 1). Number of terms of statistics taken outside

Table 20

COURSEWORK CHARACTERISTICS AND GENERALIZATION SCALES: INTERCORRELATIONS**

Coursework Characteristics	Problem 1					Problem 2				
	1	2	3	4	5	1	2	3	4	5
Trms stat-psych	33[b]	-02	13	09	01	19	-01	10	00	04
Trms stat-other	06	15	-16	-23[a]	-25[a]	08	-03	-17	-01	-28[a]
Trms stat-total	25[a]	16	-01	-13	-16	19	-06	-09	00	-18
Yrs since stat	-06	13	18	-10	11	-08	-12	03	05	01
Trms undgr stat	02	03	-13	-20	-08	05	-07	-03	-24[a]	02
Res meth-grad	09	22[a]	-07	-10	-09	21	00	-10	02	-13
Res meth-undgr	-09	-06	-24[a]	-12	-30[b]	14	-10	-17	-14	-25[a]
Trms math	10	-23[a]	-01	-18	-18	11	07	-19	-07	-26[a]
Trms engineer	07	00	03	-12	-09	00	07	-14	-19	-10
Trms physics	-03	-07	04	-20	-13	01	-01	-20	-03	-13
Trms chem	04	04	13	-19	01	07	-10	-01	05	-03
Trms sci-total	06	-09	03	-29[b]	-14	09	00	-18	-07	-19
Trms philo	-15	19	-16	-11	-12	-14	-04	-15	-03	-08
LOGIC*	15	12	-14	-06	-10	01	13	-21	-22[a]	-05
PHILO OF SCI*	10	20	-14	01	-12	05	04	-19	06	-11
EPIST*	-02	22[a]	-07	-05	-10	01	-12	-06	00	-11

*Dummy-coded with upper case letters designating higher numerical coding.

**Correlations are Pearson or point-biserial with decimal point omitted.

[a] $p < .05$ [b] $p < .01$

the Department of Psychology was inversely related to the preference for tentative over nontentative (Scale 5). Thus, subjects who had taken statistics courses outside the Department of Psychology were less likely to use this strategy. Number of undergraduate terms of statistics was inversely related to preference for task unqualified over task qualified items (Scale 4). Both number of undergraduate terms of research methods and number of terms of math were inversely related to the preference for tentative over nontentative items (Scale 5). Subjects who had more terms of these were less likely to prefer tentative over nontentative items. And finally, subjects who had more undergraduate terms of research methods were also less likely to prefer population general or college sophomore items over population specific ones.

The Attitude Variables and Generalization Scales

Correlations between the Attitude Scales and the Generalization Scales are presented in Table 21. Of the 50 correlations four are significant at the .05 level or better. This is slightly more than would be expected on the basis of sampling fluctuations alone.

Of the four statistically significant correlations only one replicated across problems. Attitude Question #7--the belief that most sciences make frequent use of significance tests--was associated with the preference for population general or college sophomore over population specific items.

Table 21

THE ATTITUDE SCALES AND GENERALIZATION SCALES: INTERCORRELATIONS*

	Problem 1					Problem 2				
	1	2	3	4	5	1	2	3	4	5
ATT SC 1	06	09[a]	08	02	05	-12	-11	01	01[b]	-01
ATT SC 2	12	24[a]	-05	-02	06	04	07	-12	30[b]	-15
ATT SC 3	-04	07	-13	-02	-06	-13	02	-11[b]	-06	-01[a]
# 7	-09	03	14	-06	10	-02	09	32[b]	-02	23[a]
# 9	05	13	-16	-05	-19	01	20	01	17	-08

*Pearson and phi correlations with decimal points omitted and not corrected for attenuation. Attitude Questions and Scales coded with higher numerical scores denoting agreement.

[a] $p < .05$ [b] $p < .01$

Background Variables and Error Variables

Relationships between the background variables and EV's are presented in Table 22. Some of the largest associations found were between major field and EV's 1 through 4. The association with major and EV 4 was the strongest. EV 4 indicates whether subjects used r as the primary factor in strength of relationship. The contingency table for major and EV 4 shows that ECO students and to a lesser extent, CL students were much more likely to err on this EV. The actual percentages of error were: ECO (71 percent), CL (46 percent), DEV (13 percent), I-O (10 percent), and S-P and EXP (both 0 percent). Several of the career preference variables also had modest correlations with EV 4 as did CL vs. NCL status.

Difference in frequencies of error between the six major fields was small for EV 5: CL (92 percent), ECO (86 percent), EXP (83 percent), I-O (80 percent), S-P (78 percent), and DEV (63 percent). Small to

Table 22

BACKGROUND VARIABLES AND ERROR VARIABLES:
CORRELATIONS AND ALTERNATE MEASURES
OF ASSOCIATION[2]

Background Variables	Error Variables 1	2	3	4	5
Major (eta)**	43[a]	40[a]	37	50[b]	25
Minor (eta) **	42	43	44	39	32
CL vs. ncl*	11	-01	30[b]	-31[b]	-20
Undgr major (eta)**	20	28	25	37	36
Undgr minor (eta)**	31	31	26	27	22
UNDGR MATH MAJ/MIN*	08	-07	-01	-05	09
Age	-06	02	-03	02	-07
FEMALE vs. male*	01	-10	-20	03	-02
Yrs grad study-psych	-07	-06	-04	-05	12
Yrs grad study-other	12	-10	-19	-12	-07
Research preference[1]	18	-05	10	-26[a]	-25[a]
Teaching preference[1]	15	-03	02	-31[b]	-26[a]
Clinical preference[1]	-12	04	-17	27[a]	22
Consult preference[1]	-26[a]	06	10	15	25[a]
Admin preference[1]	08	-02	06	05	-05
Math ability	-16	25[a]	14	20	09

*Dummy-coded with upper case letters designating higher numerical coding

**Significance levels come from Chi-square; eta from crosstabulation

[1]Low preference scored high

[2]Decimal points omitted

[a]$p < .05$ [b]$p < .01$

modest sized correlations were found for most of the career preferences and for CL vs. NCL status.

EV's 1, 2, and 3 were moderately associated with major, though for these three EV's the parallel correlations to CL vs. NCL status and the career preferences was substantially smaller. For all three EV's there were moderate differences in percentages of error for subjects in the different major fields. For EV 1 the percentages were: ECO (71 percent), S-P (33 percent), CL (32 percent), I-O (10 percent), and DEV and EXP (both 0 percent). For EV 2 the percentages were: ECO (85 percent), DEV (50 percent), S-P (44 percent), CL (41 percent), I-O (20 percent), EXP (0 percent). For EV 3 the percentages were: S-P (78 percent), DEV (75 percent), I-O (60 percent), ECO (57 percent), EXP (33 percent), CL (32 percent).

Both minor field and math ability had small to modest sized associations with the EV's. Most of these did not reach statistical significance.

Coursework Characteristics and Error Variables

Correlations between the coursework characteristics and EV's are presented in Table 23. Of the 80 correlations between these two sets of variables, eight are significant at the .05 level or better. This is more than would be expected on the basis of sampling fluctuations alone.

Several of the coursework variables measuring philosophy courses had small to moderate correlations with the EV's. Subjects who had taken a course in epistemology were more likely to have correctly used small N as a primary factor in judging strength of

Table 23

COURSEWORK CHARACTERISTICS AND ERROR VARIABLES: INTERCORRELATIONS**

Coursework Characteristics	Error Variables 1	2	3	4	5
Trms stat-psych	07	-07	-08	-02	02
Trms stat-other	14	-23[a]	-30[b]	01	13
Trms stat-total	12	-21	-29[b]	09	15
Yrs since stat	-10	12	10	03	03
Trms undgr stat	-02	03	-09	05	-02
Res meth-grad	18	-21	-16	-13	03
Res meth-undgr	08	-16	-13	01	19
Trms math	07	-03	20	-02	15
Trms engineer	00	06	18	-18	04
Trms physics	-11	10	25[a]	-02	15
Trms chem	-10	01	09	13	21
Trms sci-total	-04	-01	16	01	24[a]
Trms philo	10	-07	-20	32[b]	21
LOGIC*	-01	-05	-08	19	02
PHILO OF SCI*	-05	03	-11	10	14
EPIST*	-15	-02	-11	26[a]	34[b]

*Dummy-coded with upper case letters designating higher numerical coding

**Correlations are Pearson, point-biserial or phi coefficients with decimal points omitted

[a] $p < .05$ [b] $p < .01$

relationship with t (EV 5), and more likely to have correctly used r as the primary factor in judging strength of relationship when it was presented (EV 4). Thus, these subjects appeared to have a better understanding of how to judge strength of relationship.

Also, the number of terms of philosophy was modestly correlated with EV 4; the total number of terms of science was modestly correlated with EV 5; and the number of terms of physics was modestly correlated with EV 3.

There was a set of small to moderate correlations between two of the graduate level statistics variables and the EV's. Both the total number of terms of graduate level statistics and the number of graduate level statistics taken outside the Department of Psychology were negatively correlated with EV 2 and EV 3. Thus, the more terms of statistics, the more likely subjects were to fail to use low p as the primary factor in assurance judgments with t and r.

The Attitude Variables and Error Variables

Correlations between the Attitude Scales and EV's are presented in Table 24. Of these 25 correlations only one is significant at the .05 level or better. This is within what could be expected on the basis of sampling fluctuations alone.

Table 24

THE ATTITUDE SCALES AND ERROR VARIABLES: INTERCORRELATIONS*

	Error Variables				
	1	2	3	4	5
ATT SC 1	-03	07	08	11	-07
ATT SC 2	00	-14	-22	01	03
ATT SC 3	-01	-06	-12	07	18
# 7	-05	11	27[a]	-14	-02
# 9	09	-07	04	05	01

*Point-biserial correlations and phi coefficients with decimal points omitted.

[a] $p < .05$ [b] $p < .01$

Summary

A number of the background variables were related to either the Generalization Scales or the EV's. While major did not appear to be related to any of the Generalization Scales, CL vs. NCL status and minor field both were related to several of the scales. Minor field was related to three scales, though no one minor consistently generalized more or less. CL vs. NCL status was related to Scale 5, with clinical students more frequently choosing the cautious (tentative) strategy. CL and NCL students had very different preferences for research and clinical work and this difference was reflected in relationships between these career preferences and Scale 5. In general, the career preference variables were found to be redundant with major field, espcially CL vs. NCL status. This was because the career preference variables were highly associated with major field.

Major field was moderately associated with EV's 1 through 4,

and modestly associated with EV 5. However, CL vs. NCL status was only associated with three of these EV's. An examination of error percentages of subjects with different majors showed that there were differences between majors beyond the CL vs. NCL difference; however, because CL students made up about half the sample, when they did poorly on an EV this was picked up by the CL vs. NCL variable. The differences in majors on the EV's was reflected in associations with the career preferences and EV's. Again, redundancy between career preferences and major and CL vs. NCL status accounted for the parallel and consistent associations between the career preferences and the EV's.

There were several modest associations between minor field and the EV's.

Math ability was the only other background variable that appeared related to any of the Generalization Scales and EV's. It was associated with Scale 3 and EV 2. In both cases subjects with better math ability were less likely to generalize, or to make errors.

Some of the variables measuring number of graduate level statistics courses were related to the Generalization Scales and EV's. Subjects who had more courses of statistics in the Department of Psychology were less likely to generalize on one of the scales. Subjects who had taken more courses outside the Department did worse on two of the EV's and generalized more on one of the Scales. Total number of statistics courses was associated with less generalization on one of the Generalization Scales and more errors on two of the EV's.

Some of the undergraduate research, statistics, and math coursework variables were related to the Generalization Scales, but

not to the EV's. In general, the greater number of terms was associated with less generalization.

A number of philosophy and science coursework variables were related to the EV's. Subjects who had taken more philosophy courses, especially a course in epistemology, were more likely to have a better understanding of how to judge strength of relationship. Subjects with more total terms of science and more terms of physics were less likely to err on EV's 5 and 3, respectively. None of the philosophy or science variables appeared to be associated with the Generalization Scales.

With one exception, none of the attitude variables appeared to be related to the Generalization Scales or EV's. Attitude Question #7 was associated with a greater tendency to generalize on Scale 3.

V. DISCUSSION

A Summary of Key Findings

(1) The attitude questions showed that a large number of subjects believed that psychologists do shoddy research, though there was considerably less agreement about the belief that psychologists tend to reach the conclusions they want to even if these are not fully supported by the data. And, only a minority believed that significance tests provide objectivity.

(2) There was a fair amount of diversity in patterns of responses to the problems.

(3) Many subjects used response strategies that generalized widely on the Population, Task, and Theoretical Generality Factors. Averaging the two problems, 61 percent of the sample believed one or more population general conclusions to be valid and 14 percent of the sample believed at least half of the population general conclusions to be valid. An average of 68 percent of the sample believed one or more college sophomore conclusions to be valid and 14 percent believed at least half of these to be valid.

An average of 66 percent of the sample believed one or more task unqualified conclusions to be valid and 11 percent believed at least half of these to be valid. However, only .5 percent actually showed a preference for these over task specific conclusions.

An average of 60 percent believed one or more theoretical conclusions to be valid and 12 percent believed at least half of these

to be valid. An average of 11 percent actually showed a preference for theoretical over operational conclusions.

(4) All subjects showed a strong preference for tentative conclusions over nontentative ones.

Also, an average of 36 percent of the subjects showed a preference for tentative qualified over tentative conclusions and over 90 percent believed one or more tentative qualified conclusions to be valid. Tentative qualified conclusions reflect the misconception that the p value gives the absolute probability that the results of the experiment were due to chance.

(5) Many of the subjects did not understand the relationship of the p value and sample size to judgments of personal assurance that the null hypothesis is false, and judgements of strength of association. The most frequent error was in judgments of strength of relationship with t; 84 percent of the sample did not give evidence of knowing that a smaller N requires a substantially greater effect to reach statistical significance. Such an error illustrates the "confusion about the relationship between N, p and strength of association" described above.

The next two most frequent errors in Problems 3 through 6 were in judgments of personal assurance with t and r. About 40 percent and 48 percent of the sample, respectively, failed to use low p as the primary factor in Problems 3 and 5.

The least frequent error was the large N fallacy. Only 27 percent used large N as a primary factor in judgments of personal assurance that the null hypothesis is false.

In fact, about half the sample did not differentiate in their response strategies between judgments of personal assurance that the null hypothesis is false and judgments of strength of relationship; and 56 percent did not differentiate between the two statistics, t and r.

Thus, it appears that a substantial number of subjects do not adequately understand the concepts of strength of association, and personal assurance that the null hypothesis is false, especially in relation to sample size and the p value.

(6) The five Generalization Scales were, with one exception, unrelated to each other. Thus, there was no one group of subjects that consistently generalized more or less than others.

The five Error Variables of Problems 3 through 6 showed a pattern of relationships that seemed to largely reflect the fact that several overlapped in measuring problems, and that subjects tended to use one erroneous strategy for several different problems.

(7) There was a marked amount of inconsistency in response strategies to Problems 1 and 2. Subjects were moderately consistent in the use of the Format or Tentativity Factor, but substantially less consistent in the use of the Theoretical, Population, and Task Generality Factors.

It is likely that much of this marked inconsistency is the result of a random error process in subjects' responses to these problems. However, it appears that some of this inconsistency is the result of perceived differences in the two problems. This latter interpretation is supported by the fact that of the 28 subjects who were

inconsistent in the use of the Theoretical Generality Factor, 24 used it as a factor in Problem 1 but not in Problem 2.

Variation in the use of the Theoretical Generality Factor by some subjects may be a result of a greater similarity in meaning between operational and theoretical hypotheses in Problem 2 than in Problem 1. The operational hypothesis in Problem 2--that the left visual field is more accurate on spatial tasks--is not very different from the theoretical hypothesis--that the right hemisphere is specialized to process spatial information--especially when the accompanying assumption relating left visual field to right cerebral hemisphere is considered. The operational hypothesis for Problem 1--that it takes longer to verify complex syntactic syllogisms--is quite different from the theoretical hypothesis--that syllogisms are cognitively represented in concrete, verbal form. Thus, subjects may have been more hesitant about generalizing from the operational to the theoretical hypothesis in Problem 1 than in Problem 2.

Semantic or logical differences cannot explain all of the inconsistency found in Problems 1 and 2 because a fair amount of inconsistency was found in the use of the Format Factor, and there were no possible semantic or logical differences in this factor.

There are two explanations for the randomness in responding to Problems 1 and 2. One, the randomness was the result of subjects' uncertainty or lack of knowledge about the process of making generalizations on the basis of results from a single experiment. And two, items in the two problems reflected a way of thinking about significance tests and generalization that was foreign to subjects and thus did not

tap into their conceptions or knowledge of the area. In either case, subjects may have guessed without a stable commitment to a particular position. However, there is no way to know to what extent randomness in responding was a result of the nature of the questionnaire itself, or was the result of a lack of knowledge of subjects.

(8) Several of the background and coursework variables accounted for small amounts of the variance in generalizations and errors. The attitude variables were, with one exception, unrelated to the generalizations and errors. The randomness found in Problems 1 and 2 implies that the relationships obtained between the background, coursework and attitude variables, and the generalizations and errors will necessarily be small, and less than the true amount of association. Even if corrected for the attenuation caused by this randomness the true associations between these variables would still probably be small.

One Unresolved Question and Possible Explanations

The findings of this study raise one unresolved question: Why is it that the background, coursework, and attitude variables are not more highly correlated with the considerable individual differences in generalizations and errors? A corollary question is: What is the source of the considerable individual differences observed?

One explanation for the individual differences is that the problems presented issues or concepts that were unfamiliar or foreign to subjects. Subjects may not have perceived, understood, or even thought relevant some of the distinctions between the different levels of the four dimensions of generalizations. In Problems 3 throught 6, they may have found unfamiliar some or all of the judgments of strength

of association and personal assurance that the null hypothesis is false, and therefore were forced to guess in their responses.

It is possible, then, that subjects may not have known what to do because the concepts of significance tests and generalization represented in these problems were foreign, and unrelated to whatever knowledge they may have acquired about these procedures. However, reacting to the generalizations in Problems 1 and 2 differs little from evaluating generalizations while reading journal articles, casting some doubt on foreignness as an explanation of individual differences for the first two problems.

Another explanation for the individual differences is that the coursework variables only measure number of courses, whereas a more explanatory set of variables might measure what was actually learned in these courses. For example, it may be that a variable measuring whether subjects learned about the relationship between sample size, the p value, and strength of association may be more explanatory than just the number of statistics courses taken. Or, whether subjects learned about the empirical basis for generalization may be more important than number of courses in research methods.

A third possible explanation of individual differences in generalization in Problems 1 and 2 is that they were the result of different interpretations of the instructions in these problems. Some subjects may have assumed the instruction--to base judgments of validity only on the design and results of the experiment--to mean that they should approach the task in the same way they would evaluate any research report, following the customary practices in psychological research. Others may have taken the instructions to mean that absolutely

no external information outside of what was specifically stated should be assumed, thereby eliminating the auxiliary empirical knowledge that would customarily be used in making generalizations. Since the wording of instructions to the initial parts of Problems 3 through 6 (where subjects were asked to choose between two statistical results) was straightforward and technical, it was not likely to have been given different interpretations. Subjects may have had only a vague understanding of technical concepts involved (i.e., strength of relationship) and thus may not have known quite how to respond to these problems, but that would not have been the result of different interpretations of the instructions. Thus, this explanation would not likely account for individual differences in Problems 3 through 6.

Some Implications and Conjectures

If it is assumed that the problems in the questionnaire measured issues relevant to the process of making rational inferences and generalizations, then the findings of this study have some implications for graduate training and possibly research practices as well.

One implication is that graduate students are not receiving adequate training in aspects of the significance test methodology. It may be that the extensive training in significance test methodology fails to adequately consider the limited information a significance test really provides, and does not focus enough on the relationship between the p value, sample size, and strength of association.

The considerable individual differences in generalizations suggest that graduate students are not receiving a standardized experience or training in generalization and theory corroboration. It may be that

methodology courses concentrate so much on threats to internal validity that they ignore issues concerning generalization and theory corroboration. And it is possible that what students are learning about these issues is picked up in bits and pieces by doing theses and reading journals.

Moreover, it is possible that the heavy reliance on the significance test methodology in psychology is obscuring, perhaps even perverting, the real process of scientific inference and generalization. It may be that the emphasis on statistical significance tests has focused attention away from more basic issues such as the nature of theory corroboration and generalization, and the problem of sampling error.

If this is the case then the statistical significance test methodology should be deemphasized. The weakening of the significance test as a crutch for making scientific inferences may help motivate psychologists to pay more attention to the basic problem of making rational inferences from quantitative, statistical evidence.

In place of the heavy emphasis on statistical significance tests, graduate students should be taught a more quantitative approach to data analysis. Such an approach would emphasize measures of association, confidence intervals (as suggested by Hunter, 1979), and the theory testing approach described by Meehl (1978). They should receive coursework that deals specifically with the nature of scientific inference, generalization, theory corroboration, and the problem of sampling error.

Directions for Future Research

One direction for future research into psychologists' understanding of the statistical significance test methodology would be the further development of the questionnaire. Direct questions might be included about the meaning and use of statistical significance tests and the nature of sampling error; about sample size and Type I and Type II error; and about the nature of induction and generalization. A future questionnaire might also include some direct questions concerning the limitations of a single study (especially one with a small sample) in corroborating substantive theories and in generalizing to a population. These could be compared with responses to the six problems to see how information obtained in a more concrete situation compared with that obtained from explicit questions. Direct questions might also provide some clarification of the source of individual differences in generalization and errors. Also, the part of Problems 3 through 6 which asks about factors in subjects choice of statistical result, proved difficult for many subjects and should either be simplified or omitted.

A second goal of future research would be to determine the extent to which the generalizations and errors in the questionnaire are related to generalizations and errors in actual scientific practice, e.g., in journal articles, dissertations and theses.

Once a questionnaire to measure knowledge of the significance test methodology and generalization was perfected, one direction of future research might be to develop and evaluate educational interventions whose goal it was to improve this knowledge.

VI. SUMMARY

Mounting criticism of statistical significance tests in psychology has raised the question of how well psychologists understand the limitations of this methodology. To answer this question, the present study investigated the types of inferences psychologists believe are valid based on statistically significant results from a single experiment, and whether psychologists understand the relationship between sample size, the p value, and strength of association. The study also assesses a number of educational and attitudinal variables, which were thought to be relevant to individual differences in beliefs and understanding.

A questionnaire entitled "Conceptions of Statistics" was specially developed. It consists of three main parts. The first part includes questions on educational background and career interests; coursework in statistics, research methods, the physical sciences and mathematics, and philosophy; and ten questions designed to assess attitudes towards psychologists as researchers, psychology as a science, and significance tests. The second part included two problems each presenting a synopsis of an experiment in psychology along with its statistical results. Each synopsis was followed by 36 possible conclusions based on a completely crossed set of four factors: Tentativity Factor, Theoretical Generality Factor, Population Generality Factor, and Task Generality Factor. The same factorial structure was used for conclusions in both problems, though the specific content was different

for each. Subjects were asked to judge whether each conclusion was valid or invalid.

The third part of the questionnaire contained four problems in which subjects were asked to compare two statistical results, each of which included a sample size, a statistic value, and a significance level. Sample size and p values were systematically varied. Subjects were asked to select the result which would give them greater personal assurance that the null hypothesis was false, or the result which manifested a stronger association between independent and dependent variables. Half of the problems used t-test scores and the other half correlation coefficients.

A sample of 80 graduate students in psychology at Michigan State University was randomly selected from the 160 students enrolled for courses in Spring term, 1979.

Responses to the attitude questions showed that a large number of subjects agreed that psychologists do shoddy research and disagreed that significance tests provide objectivity.

Pattern analyses were used to classify responses to the problems, and typologies were developed from the pattern analyses. These showed considerable individual differences in responses to the problems. A moderately strong tendency for subjects to generalize widely on Population, Task and Theoretical Generality Factors was observed. However, the four generalization factors were uncorrelated. No one group of subjects consistently generalized more or less than others. Also, a majority of subjects made errors in judgments of personal assurance that the null hypothesis was false and strength of association. It

appeared that many did not understand the relationship between sample size, the p value, and strength of association.

Some of the educational, career interest, and coursework variables were associated with the generalizations and errors, but no systematic pattern of relationships was observed. None of the attitude questions were related to generalizations and errors.

For the first two parallel problems, a comparison of pattern analyses, typologies and generalizations showed there was a marked inconsistency across problems. This was partially attributed to a random error process and partially to perceived differences in content. The randomness implied that the true associations between these educational, coursework, and attitude variables, and the generalizations and errors, were larger than what was observed.

Finally, some explanations for the individual differences in generalizations and errors were given, some implications for graduate education were noted, and some directions for future research were suggested.

LIST OF REFERENCES

REFERENCES

Bakan, D. The test of significance in psychological research. In D. E. Morrison and R. E. Henkel (Eds.), The significance test controversy. Chicago: Aldine, 1970, 231-251. (Reprinted from D. Bakan, On method. San Francisco: Jossey-Bass, 1967, 1-29.

Berkson, J. Tests of significance considered as evidence. In D. E. Morrison and R. E. Henkel (Eds.), The significance test controversy. Chicago: Aldine, 1970, 285-294. (Reprinted from the Journal of the American Statistical Association, 1942, 37, 325-335.)

Bolles, R. C. The difference between statistical hypotheses and scientific hypotheses. Psychological Reports, 1962, 11, 639-645.

Carroll, D. W. A chronometric analysis of logical inference. Unpublished Ph.D. dissertation, Michigan State University, 1976.

Carver, R. P. The case against statistical significance testing. Harvard Educational Review, 1978, 48 (3), 378-399.

Fisher, R. A. The design of experiments (Fifth Edition). London: Oliver and Boyd, 1949.

Fisher, R. A. Statistical methods and scientific induction. Journal of the Royal Statistical Society, 1955, Series B, 17, 69-78.

Hays, W. L. Statistics. New York: Holt, Rinehart and Winston, 1963.

Hazelett, W. M. A criticism of statistical significance testing methodology in psychology with suggestions for an alternative approach to a theory of scientific methodological rationality based in part on a psychological theory of knowledge. Unpublished Ph.D. dissertation, University of Minnesota, 1975. Printed on demand by University Microfilms, Ann Arbor, Michigan.

Hogben, L. T. Statistical theory: The relationship of probability, credibility, and error. New York: W. W. Norton, 1957. Chapters 1-3 reprinted in D. E. Morrison and R. E. Henkel (Eds.), The significance test controversy. Chicago: Aldine, 1970, 8-56.

Hunter, J. E. Cumulating results across studies: A critique of factor analysis, canonical correlation, manova, and statistical significance testing. Invited address, American Psychological Association convention, New York, September, 1979.

Hunter, J. E. and Cohen, S. H. PACKAGE: A system of computer routines for the analysis of correlational data. Educational and Psychological Measurement, 1969, 29, 697-700.

Hunter, J. E. and Gerbing, D. W. Unidimensional measurement and confirmatory factor analysis. Occasional Paper No. 20, The Institute for Research on Teaching, Michigan State University, 1979.

Kim, J. Factor analysis. In N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Bent, Statistical package for the social sciences (Second Edition). New York: McGraw-Hill, 1975, 468-514.

Kimura, D. Spatial localization in left and right visual fields. Canadian Journal of Psychology, 1969, 23, 445-458.

Meehl, P. E. Theory testing in psychology and physics: A methodological paradox. In D. E. Morrison and R. E. Henkel (Eds.), The significance test controversy. Chicago: Aldine, 1970, 252-266. (Reprinted from Philosophy of Science, 1967, 34, 103-115.)

Neufeld, R. W. J. Clinical quantitative methods. New York: Grune and Stratton, 1977.

Nunnally, J. The place of statistics in psychology. Educational and Psychological Measurement, 1960, 20, 641-650.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Oakes, W. F. On the alleged falsity of the null hypothesis. The Psychological Record, 1975, 25, 265-272.

Popper, K. R. The logic of scientific discovery. New York: Harper and Row, 1959.

Popper, K. R. Conjectures and refutations. New York: Harper and Row, 1963.

Rosenthal, R. and Gaito, J. The interpretation of levels of significance of psychological researchers. The Journal of Psychology, 1963, 55, 33-38.

Stephenson, W. The study of behavior. Chicago: University of Chicago, 1953.

Wagner, N. M. Hand differences in braille letter learning in sighted children and adults. Unpublished M.A. thesis, Michigan State University, 1976.

Wald, A. Statistical decision functions. New York: Wiley, 1950.

Wilson, K. V. Subjectivist statistics for the current crisis. Contemporary Psychology, 1961, 6, 229-231.

APPENDIX A

The Conceptions of Statistics Questionnaire

Conceptions of Statistics

Questionnaire

General Instructions:

Following a biographical section, the questionnaire contains six numbered problems, each with several items or parts. The problems are numbered 1 through 6, though in most questionnaires exact numerical order will not be followed.

Please work on only one problem at a time. Once you have completed an entire problem and have moved on to the next one, do not go back. Also, do not look ahead to problems you have not worked on.

If you have any questions at all please ask the experimenter. Thank you very much for your participation.

Biographical Information

1. Name_____ 2. Age_____ 3. Sex _____

4. Major Field _____ 5. Minor Field _____

6. What year of graduate work in psychology is this? _____

7. Number of years of graduate work in other sciences or social sciences: _____(Which other science? _____)

8. Undergraduate major _____ 9. Undergraduate minor _____

10-14. Rank order the following five types of professional activities according to your current career interests:

research.....____
teaching.....____
clinical.....____
consulting.....____
administrative..____

15. How would you rate your mathematical ability? (Check one)

poor.....____
fair.....____
good.....____
very good.....____
excellent.....____

Coursework (For courses taken under the semester system use the following conversion formula: 1 semester = 1½ terms)

16-18. No. of graduate level terms of statistics: Dept. of Psychology ___
Other department ___
(Which?)
Total number ___

19. No. of years since your last statistics course: _____

20. No. of undergraduate terms of statistics: _____

21-22. No. of terms of research methods: graduate level _____
undergraduate _____

23-27. No. of terms (graduate + undergraduate) in: math _____
engineering _____
physics _____
chemistry _____
total _____

29. No. of terms (graduate + undergraduate) of philosophy? _____
In which of the following areas have you taken courses:
Logic _____
Philosophy of science _____
Epistemology _____

Attitude Questions

33-42. For each of the statements below place a check in the appropriate column to indicate agreement or disagreement. Do not skip any statements.

AGREE	DISAGREE	
____	____	The use of statistical significance tests prevents researchers from drawing unwarranted conclusions from their data.
____	____	Psychological researchers are not cautious enough in interpretation of their data.
____	____	The use of significance tests prevents the subjective or personal element from entering into the formulation of conclusions in psychological research.
____	____	If all studies in psychology were replicated many times, few would be consistently in agreement with each other.
____	____	Statistical tests can easily be used to show support for one's own theories, regardless of the truth.
____	____	Tests of statistical significance are the only way psychological researchers can objectively evaluate hypotheses.
____	____	Most sciences make frequent use of statistical significance tests.
____	____	Psychologists are often careless in their use of statistics.
____	____	Psychology is not as rigorous a science as physics or biology.
____	____	Psychologists are so busy trying to get publications that the quality of their research is sacrificed in the process.

Time Now _____

Problem 1

A recent study investigated the time needed to verify syllogisms (logical inferences). One of the independent variables used in the research was the form of the statements in the syllogism. The statements in each syllogism were either presented all in the affirmative form ("All A are B"), or all in the logically equivalent double negative form ("No A is not B"). The investigator was interested in the effect of syllogisms with the syntactically more complex, double negative form on verification time (time needed to determine the truth or falsity of the syllogism).

The experimenter was interested in syntactic complexity because of a theoretical interest in the nature of the internal (cognitive) representation of inferences. He reasoned that if syllogisms were stored mentally in an abstract format, the syntactic complexity of the premises would not affect verification time. However, he believed that this was not the case and predicted that verification time would be greater with the more syntactically complex syllogisms.

For his subjects, the experimenter used 50 college sophomores who were enrolled in an undergraduate psychology course and who participated in the study for special credits. Assume that the study used a design which randomly assigned subjects to one of two conditions. In one condition subjects were given syllogisms with syntactically simple (affirmative) statements. In the other condition subjects were given syllogisms with syntactically complex (double negative) statements. The mean scores were tabulated and a t-test run with the following results:

Mean Verification Time		t-test
Complex syntactic syllogisms	8.51 (n = 25)	t = 1.76
Simple syntactic syllogisms	6.27 (n = 25)	p = .04 (one-tailed)

Basing your judgments on the design of the study and its results, and only on these, indicate which of the following conclusion(s) is (are) valid. Place a check in the appropriate column for each statement below. Do not skip any statements.

VALID	INVALID	
____	____	1. It takes longer for our population to verify complex syntactic syllogisms (compared to simple syntactic ones) of the type used in this experiment.
____	____	2. We tentatively assert that it takes longer for our population to verify complex syntactic syllogisms of the type used in this experiment.

Problem 1

VALID	INVALID		
____	____	3.	We tentatively assert that it takes longer for our population to verify complex syntactic syllogisms of the type used in this experiment, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	4.	It takes longer for our population to verify complex syntactic syllogisms.
____	____	5.	We tentatively assert that it takes longer for our population to verify complex syntactic syllogisms.
____	____	6.	We tentatively assert that it takes longer for our population to verify complex syntactic syllogisms, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	7.	It takes longer for American college sophomores to verify complex syntactic syllogisms of the type used in this experiment.
____	____	8.	We tentatively assert that it takes longer for American college sophomores to verify complex syntactic syllogisms of the type used in this experiment.
____	____	9.	We tentatively assert that it takes longer for American college sophomores to verify complex syntactic syllogisms of the type used in this experiment, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	10.	It takes longer for American college sophomores to verify complex syntactic syllogisms.
____	____	11.	We tentatively assert that it takes longer for American college sophomores to verify complex syntactic syllogisms.
____	____	12.	We tentatively assert that it takes longer for American college sophomores to verify complex syntactic syllogisms, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	13.	It takes longer to verify complex syntactic syllogisms of the type used in this experiment.

Problem 1

VALID	INVALID		
____	____	14.	We tentatively assert that it takes longer to verify complex syntactic syllogisms of the type used in this experiment.
____	____	15.	We tentatively assert that it takes longer to verify complex syntactic syllogisms of the type used in this experiment, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	16.	It takes longer to verify complex syntactic syllogisms.
____	____	17.	We tentatively assert that it takes longer to verify complex syntactic syllogisms.
____	____	18.	We tentatively assert that it takes longer to verify complex syntactic syllogisms, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	19.	Syllogisms of the type used in this experiment are cognitively represented in concrete, verbal form (not abstract, symbolic form) by our population.
____	____	20.	We tentatively assert that syllogisms of the type used in this experiment are cognitively represented in concrete, verbal form by our population.
____	____	21.	We tentatively assert that syllogisms of the type used in this experiment are cognitively represented in concrete, verbal form by our population, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	22.	Syllogisms are cognitively represented in concrete, verbal form by our population.
____	____	23.	We tentatively assert that syllogisms are cognitively represented in concrete, verbal form by our population.
____	____	24.	We tentatively assert that syllogisms are cognitively represented in concrete, verbal form by our population, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	25.	Syllogisms of the type used in this experiment are cognitively represented in concrete, verbal form by American college sophomores.

Problem 1

VALID	INVALID		
____	____	26.	We tentatively assert that syllogisms of the type used in this experiment are cognitively represented in concrete, verbal form by American college sophomores.
____	____	27.	We tentatively assert that syllogisms of the type used in this experiment are cognitively represented in concrete, verbal form by American college sophomores, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	28.	Syllogisms are cognitively represented in concrete, verbal form by American college sophomores.
____	____	29.	We tentatively assert that syllogisms are cognitively represented in concrete, verbal form by American college sophomores.
____	____	30.	We tentatively assert that syllogisms are cognitively represented in concrete, verbal form by American college sophomores, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	31.	Syllogisms of the type used in this experiment are cognitively represented in concrete, verbal form.
____	____	32.	We tentatively assert that syllogisms of the type used in this experiment are cognitively represented in concrete, verbal form.
____	____	33.	We tentatively assert that syllogisms of the type used in this experiment are cognitively represented in concrete, verbal form, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	34.	Syllogisms are cognitively represented in concrete, verbal form.
____	____	35.	We tentatively assert that syllogisms are cognitively represented in concrete, verbal form.
____	____	36.	We tentatively assert that syllogisms are cognitively represented in concrete, verbal form, knowing there is a 4% probability we are in error and our results are due to chance.

Problem 1

VALID INVALID

____ ____ 37. None of the above statements are valid on the basis
True False of the design and results of the study.

38. Which of the above 37 conclusions is the best? ____

39. Are you familiar with this area of research? ___yes
___no

Time Now _____

Time Now _____

Problem 2

A recent study sought to investigate differences in the cognitive functioning of the two cerebral hemispheres of the brain. On the basis of neurological reports of patients with various brain injuries, the investigator theorized that the left hemisphere is specialized to process verbal information and the right hemisphere is specialized to process spatial information.

In order to test this theory the investigator drew on previous work which had demonstrated that the left cerebral hemisphere receives most of its direct sensory input from the right side of the body, and that the right hemisphere receives most of its direct sensory input from the left side of the body. The investigator reasoned that spatial stimuli that reached only the right hemisphere (presented to only the left visual field) would be processed more accurately than spatial stimuli that reached only the left hemisphere (presented to only the right visual field).

The spatial stimuli used were dots flashed on a screen by a tachistoscope. There were forty possible dot locations scattered throughout the screen, twenty in each visual field. A tachistoscope was used in order to insure that subjects would only use one particular side of their visual field at a time. The task involved localizing where the dot had appeared by using a second screen which had all dot locations blackened in and numbered. The dependent measure was the number of correct localizations for each visual field for each subject.

The investigator was aware of the possible effects of sex and handedness. Because she was interested in the effects of subjects sex she included it as an independent variable in her design. Handedness was controlled for by using only right-handed subjects.

The subjects were college sophomores enrolled in an introductory psychology course, who participated in the experiment for special research credits. Half were male and half were female. The order of dot presentation was randomized by visual field. Each subject saw the same standard order of presentation.

An analysis of variance was performed using mean number of correct localizations for left and right visual field for each subject:

Mean Correct Localizations			Analysis of Variance	
	Left	Right		
Male (20)	10.6	8.8	Visual Field:	F = 7.674, p = .04
Female (20)	8.9	7.1	Sex:	F = 6.435, p = .05
			Field X Sex:	F = 0.653, p = .48

Problem 2

Basing your judgments on the design of the study and its results, and only on these, indicate which of the following conclusion(s) is (are) valid. Place a check in the appropriate column for each statement below. Do not skip any statements.

VALID	INVALID		
____	____	1.	Our population is more accurate using the left visual field (than the right visual field) on spatial tasks of the type used in this experiment.
____	____	2.	We tentatively assert that our population is more accurate using the left visual field on spatial tasks of the type used in this experiment.
____	____	3.	We tentatively assert that our population is more accurate using the left visual field on spatial tasks of the type used in this experiment, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	4.	Our population is more accurate using the left visual field on spatial tasks.
____	____	5.	We tentatively assert that our population is more accurate using the left visual field on spatial tasks.
____	____	6.	We tentatively assert that our population is more accurate using the left visual field on spatial tasks, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	7.	American college sophomores are more accurate using the left visual field on spatial tasks of the type used in this experiment.
____	____	8.	We tentatively assert that American college sophomores are more accurate using the left visual field on spatial tasks of the type used in this experiment.
____	____	9.	We tentatively assert that American college sophomores are more accurate using the left visual field on spatial tasks of the type used in this experiment, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	10.	American college sophomores are more accurate using the left visual field on spatial tasks.
____	____	11.	We tentatively assert that American college sophomores are more accurate using the left visual field on spatial tasks.

Problem 2

VALID	INVALID		
____	____	12.	We tentatively assert that American college sophomores are more accurate using the left visual field on spatial tasks, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	13.	The left visual field is more accurate on spatial tasks of the type used in this experiment.
____	____	14.	We tentatively assert that the left visual field is more accurate on spatial tasks of the type used in this experiment.
____	____	15.	We tentatively assert that the left visual field is more accurate on spatial tasks of the type used in this experiment, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	16.	The left visual field is more accurate on spatial tasks.
____	____	17.	We tentatively assert that the left visual field is more accurate on spatial tasks.
____	____	18.	We tentatively assert that the left visual field is more accurate on spatial tasks, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	19.	For our population the right hemisphere is specialized to process spatial information of the type used in this experiment.
____	____	20.	We tentatively assert that for our population the right hemisphere is specialized to process spatial information of the type used in this experiment.
____	____	21.	We tentatively assert that for our population the right hemisphere is specialized to process spatial information of the type used in this experiment, knowing there is a 4% probability we are in error and our results are due to chance.
____	____	22.	For our population the right hemisphere is specialized to process spatial information.
____	____	23.	We tentatively assert that for our population the right hemisphere is specialized to process spatial information.

Problem 2

VALID INVALID

____ ____ 24. We tentatively assert that for our population the right hemisphere is specialized to process spatial information, knowing there is a 4% probability we are in error and our results are due to chance.

____ ____ 25. For American college sophomores the right hemisphere is specialized to process spatial information of the type used in this experiment.

____ ____ 26. We tentatively assert that for American college sophomores the right hemisphere is specialized to process spatial information of the type used in this experiment.

____ ____ 27. We tentatively assert that for American college sophomores the right hemisphere is specialized to process spatial information of the type used in this experiment, knowing there is a 4% probability we are in error and our results are due to chance.

____ ____ 28. For American college sophomores the right hemisphere is specialized to process spatial information.

____ ____ 29. We tentatively assert that for American college sophomores the right hemisphere is specialized to process spatial information.

____ ____ 30. We tentatively assert that for American college sophomores the right hemisphere is specialized to process spatial information, knowing there is a 4% probability we are in error and our results are due to chance.

____ ____ 31. The right hemisphere is specialized to process spatial information of the type used in this experiment.

____ ____ 32. We tentatively assert that the right hemisphere is specialized to process spatial information of the type used in this experiment.

____ ____ 33. We tentatively assert that the right hemisphere is specialized to process spatial information of the type used in this experiment, knowing there is a 4% probability we are in error and our results are due to chance.

____ ____ 34. The right hemisphere is specialized to process special information.

Problem 2

VALID INVALID

____ ____ 35. We tentatively assert that the right hemisphere is specialized to process spatial information.

____ ____ 36. We tentatively assert that the right hemisphere is specialized to process spatial information, knowing there is a 4% probability we are in error and our results are due to chance.

____ ____ 37. None of the above statements are valid on the basis
True False of the design and results of the study.

38. Which of the above 37 conclusions is the best? ____

39. Are you familiar with this area of research? ____yes
____no

Time Now _____

Problem 3

Part I. (A) n = 50, t = 1.68, p = .05

(B) n = 400, t = 2.33, p = .01

1. Above are the results of two statistical tests. Which of the two would give you greater personal assurance that your null hypothesis was false? _____

2. Which factors, both positive and negative, figured into your choice of A or B in #1 above? We want to know which of the comparisons of n, t, and p weighed in favor of your choice of A or B, and their degree of importance. We also want to know which comparisons, if any, weighed against your choice, but were only of secondary importance in your decision.

   Here are the complete headings for the five columns in the table below. Notice that they are grouped into three sections. For each of the three comparisons in the table place a check in the appropriate column.

POSITIVE FACTORS	1.	THE MOST IMPORTANT FACTOR OR CONJUNCTION OF FACTORS SUPPORTING YOUR CHOICE ABOVE
	2.	THE SECOND MOST IMPORTANT FACTOR(S) SUPPORTING YOUR CHOICE, IF ANY
	3.	THE THIRD MOST IMPORTANT FACTOR SUPPORTING YOUR CHOICE, IF ANY
NOT FACTORS	4.	COMPARISONS THAT WERE NOT FACTORS AT ALL IN YOUR CHOICE, IF ANY
NEGATIVE FACTORS	5.	COMPARISONS THAT WEIGHED AGAINST YOUR CHOICE ABOVE, IF ANY

1 MOST IMPORTANT FACTOR(S)	2 SECOND MOST IMPORTANT	3 THIRD MOST IMPORTANT	4 NOT A FACTOR	NEGATIVE FACTOR(S)	COMPARISONS
					Difference in n's
					Difference in t's
					Difference in p's

Problem 3

Part II. (A) n = 400, t = 1.65, p = .05

(B) n = 50, t = 2.41, p = .01

1. Above are the results of two statistical tests. Which of the two would give you greater personal assurance that your null hypothesis was false? _____

2. Which factors, both positive and negative, figured into your choice of A or B in #1 above? We want to know which of the comparisons of n, t, and p weighed in favor of your choice of A or B, and their degree of importance. We also want to know which comparisons, if any, weighed against your choice, but were only of secondary importance in your decision.

   Here are the complete headings for the five columns in the table below. Notice that they are grouped into three sections. For each of the three comparisons in the table place a check in the appropriate column.

	1. THE MOST IMPORTANT FACTOR OR CONJUNCTION OF FACTORS SUPPORTING YOUR CHOICE ABOVE
POSITIVE FACTORS	2. THE SECOND MOST IMPORTANT FACTOR(S) SUPPORTING YOUR CHOICE, IF ANY
	3. THE THIRD MOST IMPORTANT FACTOR SUPPORTING YOUR CHOICE, IF ANY
NOT FACTORS	4. COMPARISONS THAT WERE NOT FACTORS AT ALL IN YOUR CHOICE, IF ANY
NEGATIVE FACTORS	5. COMPARISONS THAT WEIGHED AGAINST YOUR CHOICE ABOVE, IF ANY

1 MOST IMPORTANT FACTOR(S)	2 SECOND MOST IMPORTANT	3 THIRD MOST IMPORTANT	4 NOT A FACTOR	5 NEGATIVE FACTOR(S)	COMPARISONS
					Difference in n's
					Difference in t's
					Difference in p's

Problem 4

Part I. (A) n = 50, t = 1.68, p = .05

(B) n = 400, t = 2.33, p = .01

1. Above are the results of two statistical tests. Which of the two indicates a stronger underlying relationship between the independent and dependent variables? _____

2. Which factors, both positive and negative, figured into your choice of A or B in #1 above? We want to know which of the comparisons of n, t, and p weighed in favor of your choice of A or B, and their degree of importance. We also want to know which comparisons, if any, weighed against your choice, but were only of secondary importance in your decision.

   Here are the complete headings for the five columns in the table below. Notice that they are grouped into three sections. For each of the three comparisons in the table place a check in the appropriate column.

POSITIVE FACTORS	1.	THE MOST IMPORTANT FACTOR OR CONJUNCTION OF FACTORS SUPPORTING YOUR CHOICE ABOVE
	2.	THE SECOND MOST IMPORTANT FACTOR(S) SUPPORTING YOUR CHOICE, IF ANY
	3.	THE THIRD MOST IMPORTANT FACTOR SUPPORTING YOUR CHOICE, IF ANY
NOT FACTORS	4.	COMPARISONS THAT WERE NOT FACTORS AT ALL IN YOUR CHOICE, IF ANY
NEGATIVE FACTORS	5.	COMPARISONS THAT WEIGHED AGAINST YOUR CHOICE ABOVE, IF ANY

1 MOST IMPORTANT FACTOR(S)	2 SECOND MOST IMPORTANT	3 THIRD MOST IMPORTANT	4 NOT A FACTOR	5 NEGATIVE FACTOR(S)	COMPARISONS
					Difference in n's
					Difference in t's
					Difference in p's

Problem 4

Part II. (A) n = 400, t = 1.65, p = .05

(B) n = 50, t = 2.41, p = .01

1. Above are the results of two statistical tests. Which of the two indicates a stronger underlying relationship between the independent and dependent variables? _____

2. Which factors , both positive and negative, figured into your choice of A or B in #1 above? We want to know which of the comparisons of n, t, and p weighed in favor of your choice of A or B, and their degree of importance. We also want to know which comparisons, if any, weighed against your choice, but were only of secondary importance in your decision.

   Here are the complete headings for the five columns in the table below. Notice that they are grouped into three sections. For each of the three comparisons in the table place a check in the appropriate column.

POSITIVE FACTORS	1.	THE MOST IMPORTANT FACTOR OR CONJUNCTION OF FACTORS SUPPORTING YOUR CHOICE ABOVE
	2.	THE SECOND MOST IMPORTANT FACTOR(S) SUPPORTING YOUR CHOICE, IF ANY
	3.	THE THIRD MOST IMPORTANT FACTOR SUPPORTING YOUR CHOICE, IF ANY
NOT FACTORS	4.	COMPARISONS THAT WERE NOT FACTORS AT ALL IN YOUR CHOICE, IF ANY
NEGATIVE FACTORS	5.	COMPARISONS THAT WEIGHED AGAINST YOUR CHOICE ABOVE, IF ANY

1 MOST IMPORTANT FACTOR(S)	2 SECOND MOST IMPORTANT	3 THIRD MOST IMPORTANT	4 NOT A FACTOR	5 NEGATIVE FACTOR(S)	COMPARISONS
					Difference in n's
					Difference in t's
					Difference in p's

Problem 5

Part I. (A) n = 50, r = .27, p = .05

(B) n = 400, r = .13, p = .01

1. Above are the results of two statistical tests. Which of the two would give you greater personal assurance that your null hypothesis was false? _____

2. Which factors, both positive and negative, figured into your choice of A or B in #1 above? We want to know which of the comparisons of n, r, and p weighed in favor of your choice of A or B, and their degree of importance. We also want to know which comparisons, if any, weighed against your choice, but were only of secondary importance in your decision.

   Here are the complete headings for the five columns in the table below. Notice that they are grouped into three sections. For each of the three comparisons in the table place a check in the appropriate column.

POSITIVE FACTORS	1.	THE MOST IMPORTANT FACTOR OR CONJUNCTION OF FACTORS SUPPORTING YOUR CHOICE ABOVE.
	2.	THE SECOND MOST IMPORTANT FACTOR(S) SUPPORTING YOUR CHOICE, IF ANY
	3.	THE THIRD MOST IMPORTANT FACTOR SUPPORTING YOUR CHOICE, IF ANY
NOT FACTORS	4.	COMPARISONS THAT WERE NOT FACTORS AT ALL IN YOUR CHOICE, IF ANY
NEGATIVE FACTORS	5.	COMPARISONS THAT WEIGHED AGAINST YOUR CHOICE ABOVE, IF ANY

1 MOST IMPORTANT FACTOR(S)	2 SECOND MOST IMPORTANT	3 THIRD MOST IMPORTANT	4 NOT A FACTOR	5 NEGATIVE FACTOR(S)	COMPARISONS
					Difference in n's
					Difference in r's
					Difference in p's

Problem 5

Part II. (A) n = 400, r = .10, p = .05

(B) n = 50, r = .35, p = .01

1. Above are the results of two statistical tests. Which of the two would give you greater personal assurance that your null hypothesis was false? _____

2. Which factors, both positive and negative, figured into your choice of A or B in #1 above? We want to know which of the comparisons of n, r, and p weighed in favor of your choice of A or B, and their degree of importance. We also want to know which comparisons, if any, weighed against your choice, but were only of secondary importance in your decision.

   Here are the complete headings for the five columns in the table below. Notice that they are grouped into three sections. For each of the three comparisons in the table place a check in the appropriate column.

POSITIVE FACTORS	1.	THE MOST IMPORTANT FACTOR OR CONJUNCTION OF FACTORS SUPPORTING YOUR CHOICE ABOVE
	2.	THE SECOND MOST IMPORTANT FACTOR(S) SUPPORTING YOUR CHOICE, IF ANY
	3.	THE THIRD MOST IMPORTANT FACTOR SUPPORTING YOUR CHOICE, IF ANY
NOT FACTORS	4.	COMPARISONS THAT WERE NOT FACTORS AT ALL IN YOUR CHOICE, IF ANY
NEGATIVE FACTORS	5.	COMPARISONS THAT WEIGHED AGAINST YOUR CHOICE ABOVE, IF ANY

1 MOST IMPORTANT FACTOR(S)	2 SECOND MOST IMPORTANT	3 THIRD MOST IMPORTANT	4 NOT A FACTOR	5 NEGATIVE FACTOR(S)	COMPARISONS
					Difference in n's
					Difference in r's
					Difference in p's

Problem 6

Part I. (A) n = 50, r = .27, p = .05

(B) n = 400, r = .13, p = .01

1. Above are the results of two statistical tests. Which of the two indicates a stronger underlying relationship between the independent and dependent variables? _____

2. Which factors, both positive and negative, figured into your choice of A or B in #1 above? We want to know which of the comparisons of n, r, and p weighed in favor of your choice in A or B, and their degree of importance. We also want to know which comparisons, if any, weighed against your choice, but were only of secondary importance in your decision.

   Here are the complete headings for the five columns in the table below. Notice that they are grouped into three sections. For each of the three comparisons in the table place a check in the appropriate column.

POSITIVE FACTORS	1.	THE MOST IMPORTANT FACTOR OR CONJUNCTION OF FACTORS SUPPORTING YOUR CHOICE ABOVE
	2.	THE SECOND MOST IMPORTANT FACTOR(S) SUPPORTING YOUR CHOICE, IF ANY
	3.	THE THIRD MOST IMPORTANT FACTOR SUPPORTING YOUR CHOICE, IF ANY
NOT FACTORS	4.	COMPARISONS THAT WERE NOT FACTORS AT ALL IN YOUR CHOICE, IF ANY
NEGATIVE FACTORS	5.	COMPARISONS THAT WEIGHED AGAINST YOUR CHOICE ABOVE, IF ANY

1 MOST IMPORTANT FACTOR(S)	2 SECOND MOST IMPORTANT	3 THIRD MOST IMPORTANT	4 NOT A FACTOR	5 NEGATIVE FACTOR(S)	COMPARISONS
					Difference in n's
					Difference in r's
					Difference in p's

Problem 6

Part II. (A) n = 400, r = .10, p = .05

(B) n = 50, r = .35, p = .01

1. Above are the results of two statistical tests. Which of the two indicates a stronger underlying relationship between the independent and dependent variables? _____

2. Which factors, both positive and negative, figured into your choice of A or B in #1 above? We want to know which of the comparisons of n, r, and p weighed in favor of your choice of A or B, and their degree of importance. We also want to know which comparisons, if any, weighed against your choice, but were only of secondary importance in your decision.

   Here are the complete headings for the five columns in the table below. Notice that they are grouped into three sections. For each of the three comparisons in the table place a check in the appropriate column.

POSITIVE FACTORS	1.	THE MOST IMPORTANT FACTOR OR CONJUNCTION OF FACTORS SUPPORTING YOUR CHOICE ABOVE
	2.	THE SECOND MOST IMPORTANT FACTOR(S) SUPPORTING YOUR CHOICE, IF ANY
	3.	THE THIRD MOST IMPORTANT FACTOR SUPPORTING YOUR CHOICE, IF ANY
NOT FACTORS	4.	COMPARISONS THAT WERE NOT FACTORS AT ALL IN YOUR CHOICE, IF ANY
NEGATIVE FACTORS	5.	COMPARISONS THAT WEIGHED AGAINST YOUR CHOICE ABOVE, IF ANY

1 MOST IMPORTANT FACTOR(S)	2 SECOND MOST IMPORTANT	3 THIRD MOST IMPORTANT	4 NOT A FACTOR	5 NEGATIVE FACTOR(S)	COMPARISONS
					Difference in n's
					Difference in r's
					Difference in p's

APPENDIX B

Profile Groups for Problems 1 and 2

TABLE 25

RESPONSES FOR PROFILE GROUPS 1 TO 3 (PROBLEM 1)*

PG No.	Sub. No.	Item Number 1	4	7	10	13	16	19	22	25	28	31	34
1	1	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	29	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	70	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	53	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 1 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	67	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	74	0 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
2	39	1 0 0	1 1 1	1 0 0	1 1 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	3	1 0 0	1 1 0	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	14	1 0 0	1 1 0	1 0 0	1 1 0	1 1 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	5	1 1 0	1 0 0	1 1 0	1 1 0	1 1 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	69	1 0 0	1 1 1	1 0 0	1 1 0	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	41	1 1 0	1 0 0	1 0 0	1 1 0	1 1 0	1 1 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	31	1 0 0	0 0 0	1 0 0	1 1 0	1 1 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	72	1 0 0	1 0 0	1 0 0	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	28	1 0 0	1 0 0	1 0 0	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	57	1 0 0	1 0 0	1 1 1	1 1 1	1 0 0	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	17	0 0 0	1 1 0	1 0 0	1 1 0	1 1 0	1 1 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	80	1 0 0	0 0 0	1 0 0	1 1 0	1 1 1	1 1 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
3	60	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 1	1 0 1	1 0 1	1 0 1	1 0 1	1 0 1
	65	0 0 0	1 1 0	1 0 0	1 0 0	1 0 0	1 0 0	1 1 1	1 1 0	1 1 1	1 1 1	1 1 0	1 1 0
	76	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 1 0	1 1 1	1 1 1	0 0 1	0 0 1	1 0 1	1 1 0
	25	0 0 0	0 0 0	1 0 0	1 0 0	0 0 0	1 0 0	1 1 1	1 1 0	1 1 0	1 1 0	1 1 0	1 1 0

*0 = Valid; 1 = Invalid

TABLE 26

THE FACTORIAL STRUCTURE OF ITEMS IN PROBLEMS 1 AND 2

	Item Number																																			
Factor	1			4			7			10			13			16			19			22			25			28			31			34		
Format	N	T	Q	N	T	Q	N	T	Q	N	T	Q	N	T	Q	N	T	Q	N	T	Q	N	T	Q	N	T	Q	N	T	Q	N	T	Q	N	T	Q
Population	S	S	S	S	S	S	C	C	C	C	C	C	U	U	U	U	U	U	S	S	S	S	S	S	C	C	C	C	C	C	U	U	U	U	U	U
Task	S	S	S	U	U	U	S	S	S	U	U	U	S	S	S	U	U	U	S	S	S	U	U	U	S	S	S	U	U	U	S	S	S	U	U	U
Theoret.	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T

Format: N - Nontentative level; T = Tentative level; Q = Tentative-Qualified level

Population: S = Population Specific level; C = College Sophomore level; U = Population Unqualified level

Task: S = Task Specific level; U = Task Unqualified level

Theoretical: O = Operational level; T = Theoretical level

TABLE 27

RESPONSES FOR PROFILE GROUPS 4 TO 7 (PROBLEM 1)*

PG No.	Sub. No.	Item Number 1	4	7	10	13	16	19	22	25	28	31	34
4	30	0 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0
	37	0 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0
	62	0 0 0	1 0 0	0 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0
	79	1 0 0	1 1 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0
	36	0 0 0	1 0 0	0 0 0	1 0 0	0 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0
	45	0 0 0	1 0 0	0 0 0	1 0 0	1 0 0	1 0 0	1 1 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0
5	58	0 0 0	1 1 0	1 0 0	1 0 0	1 0 0	1 0 0	1 0 1	1 1 1	1 0 0	1 0 0	1 0 0	1 0 0
	27	1 0 0	1 0 0	1 0 1	1 1 1	1 1 0	0 1 0	0 1 0	0 1 0	1 0 1	1 0 1	1 0 0	1 0 0
	12	1 0 0	1 0 0	1 1 0	1 1 0	1 0 0	1 1 0	1 0 0	1 1 1	1 0 0	1 1 0	1 0 0	1 0 0
	15	0 0 0	1 0 0	1 1 1	1 1 1	1 0 0	1 1 1	1 0 0	1 0 0	1 1 1	1 1 0	1 0 0	1 0 0
	71	0 0 0	1 0 0	1 0 0	1 1 1	1 0 0	1 0 0	1 0 0	1 1 0	1 1 0	1 1 0	1 0 0	1 1 1
6	10	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	11	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	20	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	75	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	40	0 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	64	0 0 0	0 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	6	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
7	50	1 0 0	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1	1 0 0	1 1 1	0 0 0	1 1 1	1 1 1	1 1 1
	46	1 0 0	1 0 0	1 0 0	1 0 0	1 1 1	1 1 1	1 0 0	1 1 1	1 0 0	1 1 1	1 0 0	1 1 1
	32	0 0 0	1 1 1	0 0 0	1 1 1	1 1 1	1 1 1	1 0 0	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1
	51	1 1 0	1 1 1	1 0 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	0 0 1	1 1 1	1 1 1	1 1 1
	2	0 0 0	1 0 0	1 0 0	1 0 0	1 1 1	1 1 1	1 0 0	1 1 1	1 0 0	1 0 0	1 0 0	1 1 1
	24	1 0 0	1 1 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 0	1 1 1	1 0 0	1 1 0	1 1 1	1 1 1

*0 = Valid; 1 = Invalid

TABLE 28

RESPONSES FOR PROFILE GROUPS 8 TO 12 (PROBLEM 1)*

PG No.	Sub. No.	Item Number 1	4	7	10	13	16	19	22	25	28	31	34
8	54	0 0 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1						1 0 0 1 0 0 1 1 1 1 1 1 1 1					
	77	1 0 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1						1 0 1 1 0 0 1 1 1 1 1 1 1 1					
	34	1 0 0 1 0 0 1 1 0 1 1 1 1 1 0 1 1 1						1 1 0 1 0 0 1 1 0 1 1 0 1 1					
	68	1 0 0 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1						0 0 0 0 0 0 1 0 0 1 1 1 1 1					
	61	1 1 1 1 0 0 1 1 0 1 1 1 1 1 1 1 1 1						1 1 1 1 1 0 1 1 1 1 1 0 1 1					
	42	1 0 0 1 0 0 1 1 1 1 1 1 1 1 1 1 0 0						1 0 0 1 0 0 1 1 1 1 0 0 1 0					
	44	0 0 0 1 0 0 1 0 0 1 0 0 1 1 1 1 1 1						1 0 0 1 0 0 1 0 0 1 0 0 1 1					
9	21	0 1 0 1 1 1 1 0 1 1 0 1 1 0 1 1 0 1						1 1 1 1 1 1 1 1 1 1 1 1 1 1					
	49	0 0 1 1 0 1 1 0 1 1 0 1 0 0 1 1 0 1						1 1 1 1 1 1 1 1 1 1 1 1 1 1					
	43	0 0 0 1 1 1 1 0 0 1 0 0 1 0 1 1 0 1						1 1 1 1 1 1 1 1 1 1 1 1 1 1					
	16	0 1 0 1 1 1 1 0 0 1 0 1 1 0 1 1 0 1						1 0 0 1 0 0 1 1 1 1 1 1 0 1					
10	63	0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1						1 1 1 1 1 1 1 1 0 1 1 1 1 1					
	19	0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1						1 1 1 1 1 1 1 1 1 1 1 1 1 1					
11	55	1 0 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1						1 1 1 1 1 1 1 1 1 1 1 1 1 1					
12	59	0 0 0 1 1 1 0 0 0 1 1 1 0 0 0 1 1 1						1 1 1 1 1 1 1 1 1 1 1 1 1 1					
	13	0 0 1 1 0 1 1 1 1 1 1 1 0 0 0 1 1 0						1 1 1 1 1 1 1 1 1 1 1 1 1 1					
	18	0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1						0 0 0 1 1 1 1 1 1 1 1 1 1 0					
	56	0 0 0 1 1 1 1 0 0 0 1 1 1 0 0 0 1 1						1 1 1 1 1 1 1 1 1 1 1 1 1 1					

*0 = Valid; 1 = Invalid

TABLE 29

RESPONSES FOR PROFILE GROUPS 13 TO 16 (PROBLEM 1)*

PG	Sub.	Item Number																																			
No.	No.	1			4			7			10			13			16			19			22			25			28			31			34		
13	35	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	22	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
14	33	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0
	52	0	1	0	1	0	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0
15	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	23	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	38	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	47	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	66	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16	48	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0
	78	0	0	0	1	1	0	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	0	1	1	1	1	1	1
	26	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	1	1	0	1	0	0	1	0	0	1	1	1	1	1	1
	7	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	1	1	0	1
	9	0	0	0	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0	0	1	1	0	1	1	1
	73	0	0	0	1	1	0	1	0	0	0	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1
	8	1	0	0	1	0	0	1	0	0	1	0	1	1	0	0	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1

*0 = Valid; 1 = Invalid

TABLE 30

RESPONSES FOR PROFILE GROUPS 1 AND 2 (PROBLEM 2)*

PG No.	Sub No.	Item Number 1	4	7	10	13	16	19	22	25	28	31	34
1	30	1	0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0	1 0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0
	37	1	0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0	1 0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0
	45	1	0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0	1 0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0
	62	1	0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0	1 0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0
	36	0	0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0	1 0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0
	44	0	0 0 1	0 0 1	0 0 1	1 0 1	0 0 1	0 0	1 0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0
	77	1	0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0	1 0 0 1	1 0 1	0 0 1	0 1 1	0 1 1	0 1	0 1
	71	0	0 0 0	0 0 0 1	0 0 1	0 0 1	0 0 1	0 0	1 0 0 1	0 0 1	0 0 1	1 0 1	1 0 1	1 0 1	1 0
	60	1	0 0 1	0 0 1	0 0 1	0 0 1	0 0 1	0 0	1 0 1 1	0 1 1	0 1 1	0 1 1	0 1 1	0 1 1	0 1
	27	1	0 0 1	0 0 1	1 1 1	1 0 1	1 0 0	1 0 0	1 0 0 1	1 1 1	1 1 1	1 1 1	0 0 1	0 0	
	15	0	0 0 1	0 0 1	1 1 1	1 1 1	0 0 1	1 0	0 1 0 0	1 0 0 1	0 0 1	1 0 1	0 0 1	0 0	
	53	1	0 0 1	0 0 1	0 0 1	1 1 1	0 0 1	1 1	0 0 1 0	0 1 0 0	1 1 1	1 0 0 1	0 0 1	0 0	
	78	0	0 0 1	0 0 1	0 0 1	0 0 1	1 0 1	1 0	1 1 1 1	0 0 0 0	0 1 0 0	1 0 0 1	0 0 1	0 0	
	26	0	0 0 1	0 0 0	0 0 0 1	0 0 0	0 0 1	0 1	1 0 0 1	0 0 1	0 0 1	0 0 1	1 1 1	1 0 0	
	16	0	0 0 1	0 0 1	1 1 1	1 1 1	0 0 1	0 0	1 0 0 1	0 0 1	1 1 1	1 1 1	1 1 0	0 1 0	0
	34	0	0 0 1	0 0 1	1 1 1	1 1 1	0 1 1	0 1	1 0 1 1	0 0 1	1 1 1	1 1 1	1 1 0	1 1 0	1
2	10	1	0 0 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1	1
	11	1	0 0 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1	1
	55	1	0 0 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1	1
	72	1	0 0 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1	1
	29	1	0 0 1	1 1 1	1 1 1	1 1 1	0 1 1	1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1	1
	63	0	0 0 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1	1
	67	0	0 0 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1	1
	28	1	0 0 1	1 1 1	1 1 1	1 1 1	0 0 1	1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1	1
	69	1	1 0 1	1 1 1	1 1 0	1 1 1	1 1 1	1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1	1

*0 = Valid; 1 = Invalid

TABLE 31

RESPONSES FOR PROFILE GROUPS 3 TO 5 (PROBLEM 2)*

PG No.	Sub No.	Item Number 1	4	7	10	13	16	19	22	25	28	31	34
3	40	0 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	66	0 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	32	0 0 0	1 1 1	0 0 0	1 1 1	1 1 1	1 1 1	1 0 0	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1
	68	0 0 0	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1	0 0 0	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1
	39	1 0 0	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1	1 0 0	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1
	50	1 0 0	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1	1 0 0	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1
	24	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
	54	0 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 0 0	1 1 1	1 1 1	1 1 1	0 0 1	1 1
	9	0 0 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 0	1 1 1	1 1 1	1 1 1	1 0 1	1 1
	20	1 1 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 0	1 1 1	1 1 1	1 1 1	1 1 1	1 1
	59	0 0 0	1 1 1	0 0 0	1 1 1	0 0 0	1 1 1	1 0 0	1 1 1	1 0 0	1 1 1	0 0 1	1 1
	12	0 0 0	1 1 1	1 1 0	1 1 1	1 1 1	0 1 1	1 0 1	1 1 1	1 0 1	1 1 1	0 1 1	1 1
4	52	1 1 0	1 1 0	1 1 0	1 1 0	1 1 0	1 1 0	1 1 0	1 1 0	1 1 0	1 1 0	1 1 0	1 1 0
	80	1 1 0	1 1 1	1 1 0	1 1 1	1 1 1	0 1 1	0 1 1	0 1 1	0 1 1	0 1 1	0 1 1	0 1 1
	31	1 1 0	1 1 0	1 1 0	1 1 1	0 1 1	1 1 1	1 1 0	1 1 0	1 1 0	1 1 0	1 1 0	1 1 0
	61	1 1 0	1 1 1	1 1 0	1 1 1	1 1 1	0 1 1	1 1 0	1 1 0	1 1 0	1 1 0	1 1 0	1 1 1
	33	1 1 0	1 1 1	1 1 0	1 1 1	1 1 1	0 1 1	1 1 0	1 1 1	1 1 0	1 1 1	1 1 0	1 1 1
	5	1 1 0	1 1 1	1 1 0	1 1 1	1 1 1	0 1 1	1 1 0	1 1 1	1 1 0	1 1 1	1 1 1	1 1 1
5	22	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 0	1 1 1	1 1 1	1 1 1
	35	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 0	1 1 1	1 1 1	1 1 1
	7	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 0	1 1 0	1 1 1	1 1 1

*0 = Valid; 1 = Invalid

TABLE 32

RESPONSES FOR PROFILE GROUPS 6, 7, 8, 9 AND 15 (PROBLEM 2)*

PG No.	Sub. No.	Item Number 1 4 7 10 13 16 19 22 25 28 31 34
6	3	1 0 0 1 0 0 1 0 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
	57	1 0 0 1 1 1 1 0 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
	43	0 0 1 1 0 0 1 0 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
	41	1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
	14	1 0 0 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
	70	0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
	2	0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1
7	48	0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
	64	0 0 0 1 1 1 1 0 0 1 1 1 0 0 0 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1
	58	0 0 0 1 1 1 1 1 0 1 1 0 0 0 0 1 1 0 1 1 0 1 1 1 1 0 0 1 1 0 0 0 0 1 0 0
8	13	1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
9	75	1 0 0 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1
	6	1 0 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1
	17	1 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1
15	19	0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
	76	0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

*0 = Valid; 1 = Invalid

TABLE 33

RESPONSES FOR PROFILE GROUPS 10 THROUGH 15 (PROBLEM 2)*

PG No.	Sub. No.	Item Number 1			4			7			10			13			16			19			22			25			28			31			34		
10	73	1	0	0	1	1	0	1	1	0	1	1	0	1	1	0	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1
	51	1	1	0	1	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
	18	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	49	0	0	1	0	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
	8	1	0	0	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1
	74	0	0	0	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1
	21	0	0	0	0	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1
12	1	1	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1
	79	1	0	0	1	0	0	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	23	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	38	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	42	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	47	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	56	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	46	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	0	0	1	0	0	1	0	0	1	1	1	1	1	1
	25	0	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	65	0	0	0	1	1	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	1	1	1

*0 = Valid; 1 = Invalid.

Profile Groups - Problem 1

PG's 1, 2, and 3 emerged together in one large group in the factor analysis. During modification of these blindly formed groups this was subdivided into three smaller, and more homogenous groups each with slightly different response patterns. The data for these three groups, named PG's 1 to 3, are presented in Table 25.

Subjects in PG 1 make a sharp distinction between the two levels of the Theoretical Generality Factor and between the two tentative levels (Tentative and Tentative-Qualified) and one nontentative level of the Format Factor. (See Table 26 for a display of the factorial structure of the items in Problems 1 and 2). They reject (judge invalid) all theoretical conclusions. For operational conclusions they reject only those that are nontentative in format. The response strategy inferred from this pattern is to accept (judge valid) only those conclusions that are operational and tentative. For this group the difference between the two tentative formats appears not to be a factor in decisions about validity.

PG 2 is highly similar to PG 1. Subjects in this group also reject all theoretical conclusions and they reject operational conclusions that are not tentative in format. The main difference between these two groups is that PG 2 shows a tendency to differentiate between the two types of tentative format. Whereas subjects in PG 1 accept all tentative, operational items, subjects in PG 2 tend to accept only tentative qualified ones, hence rejecting a number of tentative, operational items. The strategy inferred from this pattern uses two decision rules, a main rule and a secondary rule. The main rule is: item is

valid only if operational and tentative. The secondary rule is: reject some tentative items if they are not tentative qualified. This secondary rule, then, is an elaboration of the main rule. It is applied in slightly different ways by different subjects.

PG 3 is also highly similar to PG 1. For the operational conclusions an identical response pattern is evident. However, for the theoretical conclusions a slightly different pattern obtains as not all theoretical items are rejected. For this group, there is less of a distinction evident between operational and theoretical conclusions. The response strategy inferred for this group consists of a main rule and a secondary rule. The main rule is identical to that of PG 2: valid if operational and tentative. The secondary rule is: accept some tentative theoretical items as valid especially if they are tentative qualified in format. Again, this secondary rule is applied in slightly different ways by different subjects in the group.

The data for PG's 4 to 7 are presented in Table 27. For PG 4 the Theoretical Generality Factor does not appear important at all in decisions about validity of conclusions. Decisions appear to be based almost solely on the Format Factor. The rule followed by this group is to accept only tentative items. The one exception to this is item 1, a nontentative item which is also accepted; possibly it was so conservative on the other three factors that tentativity no longer seemed necessary. This suggests that at least in some slight or peripheral way the other three factors were taken into consideration by these subjects. However, in further analyses of the PG's, this

group will be considered to have only used the Format Factor.

PG 5 shows some similarity to PG 4 in that decisions were based primarily on the Format Factor. In this group there were slight differences between subjects and some apparent inconsistencies for several individuals. The rule that all subjects had in common was to reject nontentative items. Some tentative items were rejected too, but it was not clear what rule was applied in doing this. It is possible that some inconsistent use was made of Population Generality and Task Generality Factors, but for purposes of further analyses this is ignored.

Subjects in PG 6 followed a pattern in which they rejected all conclusions except the two least generalized (on Population, Task, and Theoretical Factors), tentative items. In this more conservative decision rule subjects took into account all four factors.

Responses for subjects in PG 7 show a complex pattern that seems to be based on a weighted combination of three or possibly four factors. In order of importance the three factors used were the Format Factor, the Task Generality Factor, and the Population Generality Factor. The Theoretical Generality Factor may have been given a very slight weighting in the decision, but for the sake of clarity this will be ignored. An attempt to construct a rule to account for this complex response pattern yields the following: judge conclusion valid when a favorable weighted combination of tentative format, task specific, and population specific or college sophomore type exists.

PG's 8 to 12 are presented in Table 28. Subjects in PG 8 also showed a complex pattern of response that seems to be based on a

weighted combination of three factors. The strategy constructed to account for this pattern has three rules of different degrees of importance. The most important is to accept conclusion if tentative in format, especially if item is tentative qualified. The second most important rule is to accept conclusion if it is population specific. And the least important is to accept the item if it is task unqualified.

PG 9 is similar to PG 2 in its fairly clear preference for rejecting all theoretical items and a strategy for deciding about operational items based on the Format Factor. The main difference is that subjects in PG 9 showed somewhat more of a preference for tentative over tentative qualified items whereas the opposite was true in PG 2. Also, subjects in PG 9 appeared to give some slightly greater consideration to Population and Task Factors, which were not as central in PG 2. This can be seen in the acceptance by all subjects in PG 9 of item 1, the least generalized item, and in the use of the Format Factor at different population and task levels.

The subjects of PG 10 followed a clearcut pattern in which they only accepted as valid operational items which were population specific. Thus, their decision was only based on Theoretical Generality and Population Generality Factors with the Format and Task Generalization Factors not used at all. A similar rule was followed in PG 11, but here format was an additional consideration, as items were only accepted if they were also tentative in format.

PG 12 was not a particularly homogeneous group. This may be due in part to subjects not using a consistent rule for all items. The rule common to members of this group is to accept only operational items

which are task specific, and population specific or population unqualified. However, subjects show some idiosyncratic deviations from this strategy, which might be better described as a rough guideline than an actual rule.

The data for PG's 13 to 16 is presented in Table 29. The two subjects in PG 13 accept only the tentative qualified, task specific, college sophomore item. One subject accepts both its operational and theoretical forms, while the other accepts only its operational form. For purposes of further analyses this group is considered to have based their response strategy on all four factors.

The two subjects in PG 14 both accept only tentative qualified items. However, the second subject makes some differentiation based on the Population, Theoretical, and the Task Factors. This subject is more likely to accept an item if it is population and task specific, and operational.

PG 15 is that group of subjects not included in the factor analysis and blind clustering procedures because they rejected all 36 items for Problem 1. It is probable that these subjects based their judgments not on the items themselves but on something they believed was missing from the research description and which prevented justification of any conclusion.

PG 16 is a residual group of those subjects which could not be classified into profile groups. Some of these subjects used rules that were unique and dissimilar from those of all the other groups. Some appeared to be inconsistent in the rules they followed.

Profile Groups - Problem 2

The data for PG's 1 and 2 are presented in Table 30. Subjects in PG 1 used one major rule and two minor rules. The major rule for this group is: accept all tentative items. There were, however, scattered deviations from this rule which required the formulation of the two minor rules. The first of these is to accept some nontentative items, too, at the lowest levels of generality for the Population, Task and Theoretical Generality Factors. The second minor rule is to reject tentative items at higher levels of generality for these factors. Thus, the Format Factor served as the basis for judgments of validity, though in certain extreme cases subjects wavered slightly from sole use of this factor. For purposes of further analyses, however, this PG was considered to have used only the Format Factor.

The subjects in PG 2 followed a pattern in which they judged valid only the least generalized items--those that were population specific, task specific, and operational. Most of these subjects accepted only those that were tentative in format, though two subjects accepted all of the first three items. For purposes of further analyses this entire group was considered to have used the Population, Task, and Theoretical Factors.

The data for PG's 3 to 5 are presented in Table 31. The data for PG 3 showed a complex pattern of responses. The rule constructed to account for this pattern is: accept an item when it embodies a favorable weighted combination of four conditions (in order of importance): task specific, population specific or college sophomore type, tentative, and operational. There is a certain amount of variation

among subjects evident in the data for this PG. This appears to be principally in the use of the Population Generality Factor.

The six subjects in PG 4 show a highly similar pattern, though like PG 3 some variation among them is evident. The main rule used by this group is to accept only tentative qualified conclusions. However, not all tentative qualified conclusions are accepted, suggesting the existence of a secondary, minor rule for this group. This secondary rule appears to be based on the Population and Task Generality Factors. The secondary use of these two factors is more pronounced for the fourth through sixth subjects of the group (#'s 61, 33, and 5). For this reason further analyses considered these three individuals to have used the Population and Task Factors, whereas the first three subjects were only considered to have used the Format Factor. None of the subjects appeared to use the Theoretical Generality Factor.

PG 5 followed a fairly straightforward pattern. They accepted as valid only the items that were tentative qualified, task specific, and college sophomore type. These were accepted at both levels of the Theoretical Generality Factor, suggesting that this factor was not used at all in judgments of validity.

The data for PG's 6 through 9, and 15, are presented in Table 32. PG 6 is only a moderately homogeneous group. Subjects appear to use variations of the following rule: accept as valid conclusions which are tentative in format and not very generalized on Population, Task, and Theoretical Factors.

PG 7 is a mediocre PG. The three subjects in this group follow a somewhat similar pattern, but there are significant dissimilarities between them, and there are some inconsistencies within each. The general

response strategy common to all three is a tendency to accept as valid items that are task specific, either population specific or population unqualified, and to prefer operational items. Two of the subjects seem to make distinctions among the different levels of the Format Factor, while the other does not. This differentiation was taken into account in later analyses.

PG 8 at first included three subjects (#'s 13, 19 and 76). Further examination suggested that this group should be split up and subjects 19 and 76 were placed in a new group, PG 15. This was done because it appeared that the two groups followed response strategies that were divergent enough to classify separately. The lone subject in PG 8 accepted as valid only those operational conclusions that were population specific and task unqualified. The two subjects in PG 15 accepted as valid operational conclusions that were population specific, but it was not necessary that they be task unqualified, too. One of these subjects also accepted population unqualified and task specific, operational items. None of these subjects differentiated conclusions on the basis of their tentativity.

The subjects in PG 9 followed a general rule in which they accepted as valid tentative, task specific, and population specific or population unqualified items. One subject also accepted the college sophomore items which were tentative and task specific. Thus, the one factor this group did not use was the Theoretical Factor.

The data for PG's 10 through 14 are presented in Table 33. Subjects in PG 10 seemed inconsistent in their responses patterns. They all tended to follow a general rule in which the Format Factor was the

most important determinant of their responses, with the Population and Task Factors being of somewhat lesser importance, depending on the subject. Subjects appeared to use these factors inconsistently. The Theoretical Generality Factor seemed to be of very little if any importance in response strategies.

The rule clearly followed by three of the four subjects in PG 11 is to accept as valid only tentative unqualified conclusions, though two of the subjects deviated from this for the first few items. The first subject also accepted only tentative qualified items but did so primarily for operational items. Thus, for purposes of further analyses this first subject was also considered to have used the Theoretical Factor in his decision rule.

The two subjects in PG 12 both accepted as valid tentative conclusions that were population specific. The second accepted them only if they were operational, while the first subject did not distinguish between levels of the Theoretical Generality Factor, accepting these items for both. The second subject also accepted a few other tentative, operational items.

PG 13 was that group of six subjects who judged as invalid all conclusions in Problem 2, probably basing their rejection on the nature of instructions given for the problem or on the research design and statistical results provided.

PG 14 was the residual group of subjects that could not be classified in a PG. Responses patterns for these tended to be inconsistent, odd, and difficult to characterize.

APPENDIX C

Profile Groups for Problems 3 Through 6

The Profile Groups - Problems 3 through 6

PG 1 includes 18 subjects. Choice responses for these subjects are presented in Table 34. Subjects in PG 1 chose A for Problem 6(I), and B for other comparisons. The rule inferred from this is: for statistical results including t scores subjects use as their main factor low values of p (or possibly high values of t, or possibly both); they favor high values of r for strength of relationship judgments where r is given, and they favor low values of p for personal assurance judgments. MVS's (see Table 35) clearly confirm this as well as ruling out the possible joint use of t in statistical results involving t-test scores.

Table 34

CHOICE RESPONSES FOR PG 1*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
7	1	1	1	1	1	1	0	1
13	1	1	1	1	1	1	0	1
15	1	1	1	1	1	1	0	1
21	1	1	1	1	1	1	0	1
22	1	1	1	1	1	1	0	1
31	1	1	1	1	1	1	0	1
40	1	1	1	1	1	1	0	1
42	1	1	1	1	1	1	0	1
44	1	1	1	1	1	1	0	1
57	1	1	1	1	1	1	0	1
59	1	1	1	1	1	1	0	1
61	1	1	1	1	1	1	0	1
75	1	1	1	1	1	1	0	1
77	1	1	1	1	1	1	0	1
78	1	1	1	1	1	1	0	1
80	1	1	1	1	1	1	0	1
58	0	1	1	1	1	1	0	1
38	1	1	1	0	1	1	0	1

*A = 0, B = 1

TABLE 35

WEIGHTED DIRECTIONAL PREFERENCES FOR PROBLEMS 3 THROUGH 6:
MEAN VECTOR SCORES FOR THE PG's*

Profile	3						4						5						6					
	I			II			I			II			I			II			I			II		
Group	N	t	p	N	t	p	N	t	p	N	t	p	N	r	p	N	r	p	N	r	p	N	r	p
1	13	10	29	06	12	29	10	14	27	04	12	26	15	13	27	03	19	28	-01	29	10	-01	29	13
2	-30	04	11	-30	15	24	-30	04	11	-30	15	24	-28	23	11	-28	25	22	-28	26	11	-28	26	19
3	27	12	20	30	08	15	27	10	20	30	07	14	25	08	23	28	08	18	27	08	22	30	08	15
4	13	23	23	04	21	23	13	24	22	03	20	19	-02	28	07	-01	28	16	-05	28	11	-05	29	13
5	30	05	20	30	08	15	20	05	30	15	05	30	30	08	20	30	08	15	20	08	30	15	15	25
6	28	10	16	30	06	09	20	26	04	12	28	10	28	09	20	30	09	09	06	30	00	06	30	00
7	30	00	00	30	00	00	-30	00	00	-30	00	00	30	00	20	30	00	00	-30	20	00	-30	20	10
8	25	00	25	30	00	08	30	00	25	30	00	08	25	15	25	30	15	15	08	30	08	08	30	05
9	03	10	07	00	10	27	-20	-20	00	23	-17	00	00	15	07	-20	23	27	-10	30	-02	-02	30	17
10	20	00	30	-30	00	20	00	-30	15	00	-30	15	15	30	-20	15	20	30	15	30	15	00	30	20
11	08	00	30	00	00	30	-30	00	15	-27	00	27	08	05	30	00	17	27	-05	30	00	-05	30	20
12	-10	-30	15	30	-20	15	10	20	30	15	20	30	-10	20	15	-10	20	30	10	15	30	15	20	30
13	30	13	20	30	05	15	30	13	20	30	00	15	20	20	12	02	30	17	13	20	22	03	30	17
14	03	05	30	30	08	15	03	05	30	15	-08	30	20	08	30	15	20	30	20	08	30	15	20	30
15	20	10	30	-30	20	30	20	10	30	-30	20	30	30	15	30	30	15	15	30	15	20	30	15	15
16	14	11	29	08	11	29	14	11	28	09	11	27	21	06	28	14	11	27	19	07	27	12	11	28

*Decimal points have been omitted between digits.

PG 2 contains eight subjects. Data for this group is presented in Table 36. Subjects in this group chose statistical result A for all subsection I comparisons and B for all subsection II comparisons. The decision rule inferred from this is: choose the result with the small N for all problems and possibly as a secondary factor, favor results with large r. This rule was clearly confirmed by the MVS's, which indicated that large r was used as a secondary factor in items that contained r.

Table 36

CHOICE RESPONSES FOR PG 2*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
5	0	1	0	1	0	1	0	1
18	0	1	0	1	0	1	0	1
23	0	1	0	1	0	1	0	1
56	0	1	0	1	0	1	0	1
63	0	1	0	1	0	1	0	1
66	0	1	0	1	0	1	0	1
70	0	1	0	1	0	1	0	1
36	0	1	0	0	0	1	0	1

*A = 0; B = 1

PG 3 used a pattern of choices opposite to that used by PG 2. They chose B for all I's and A for all II's. The decision rule for these subjects was to favor large N for all problems. The MVS's clearly confirm this strategy. They also indicate that subjects use low p values as a secondary factor. Data for this group is presented in Table 37.

PG 4 contains 12 subjects. Data for this group is presented in Table 38. Subjects in this group chose B for all problems using t.

Table 37

CHOICE RESPONSES FOR PG 3*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
11	1	0	1	0	1	0	1	0
19	1	0	1	0	1	0	1	0
30	1	0	1	0	1	0	1	0
37	1	0	1	0	1	0	1	0
41	1	0	1	0	1	0	1	0
27	1	0	1	0	1	1	1	0

*A = 0; B = 1

For problems using r they chose A for the two subsections I's, and B for the two subsection II's. From this it is inferred that in problems using r subjects chose the result with higher r and in problems using t subjects chose the result with higher t, or possibly with low p, too. This rule is confirmed by the MVS's. The MVS's also suggest the combined use of t and p for problems containing t-test scores.

Table 38

CHOICE RESPONSES FOR PG 4*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
10	1	1	1	1	0	1	0	1
34	1	1	1	1	0	1	0	1
35	1	1	1	1	0	1	0	1
43	1	1	1	1	0	1	0	1
45	1	1	1	1	0	1	0	1
50	1	1	1	1	0	1	0	1
51	1	1	1	1	0	1	0	1
64	1	1	1	1	0	1	0	1
65	1	1	1	1	0	1	0	1
69	1	1	1	1	0	1	0	1
72	1	1	1	1	0	1	0	1
33	1	0	1	1	0	1	0	1

*A = 0; B = 1

PG 5 is a small group with only two subjects. Data for this group is presented in Table 39. Subjects in this group picked choice B for all problems in which the judgment was strength of relationship. They also picked B for subsection I of Problems 3 and 5--which asked about personal assurance. For subsection II of these two problems they picked choice A. The rule inferred from these choices is: use low p (or possibly high t for items with t) for all strength of relationship judgments; use high N for all assurance judgements. The MVS's support the use of p and N. They suggest high N as a secondary factor to p in strength of relationship judgments, and low p as a secondary factor to N in assurance judgments. Thus, subjects use high N and low p in all problems but weight them differently depending on the type of judgment involved.

Table 39

CHOICE RESPONSES FOR PG 5*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
12	1	0	1	1	1	0	1	1
16	1	0	1	1	1	0	1	1

*A = 0; B = 1

PG 6 has five subjects. Data for this group is presented in Table 40. This group chose B for the first subsection of Problems 3 and 5, the assurance problems. They chose A for the second subsection of these problems. For strength of relationship problems they chose B if the problem had t, and if the problem had r, they chose A for subsection I and B for subsection II. From this complex pattern the

Table 40

CHOICE RESPONSES FOR PG 6*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
46	1	0	1	1	1	0	0	1
53	1	0	1	1	1	0	0	1
54	1	0	1	1	1	0	0	1
60	1	0	1	1	1	0	0	1
74	1	0	1	1	1	0	0	1

*A = 0; B = 1

following decision rule is inferred: use high N for all problems with assurance judgments; for problems with strength of relationship judgments, use high t or low p for those with t, and use r for those with r. Thus, the statistic was important for strength of relationship and sample size was important for personal assurance. This strategy was confirmed by the MVS's, which suggested that low p was not very important in strength of relationship judgements with t.

PG 7 has only one subject. Data for this subject is presented in Table 41. This subject followed a consistent pattern of choices within judgment types. For assurance judgments, B was chosen for the subsection I's, while for subsection II's, A was chosen. For strength of relationship judgments the reverse pattern was used--A for subsection I's and B for subsection II's. From this it is inferred that the subject used the following rule: large N favored for assurance and small N for strength of relationship. It was also possible that the subject used high r as a secondary factor in strength of relationship judgments which included correlations. This latter rule was in fact confirmed by the MVS's as was the main decision rule.

Table 41

CHOICE RESPONSES FOR PG 7*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
73	1	0	0	1	1	0	0	1

*A = 0; B = 1

Data for PG 8 is presented in Table 42. The two subjects in this group followed the same choice pattern for Problems 3 through 5. They chose B for subsection I's and A for subsection II's. For Problem 6 they reversed their pattern, choosing A for subsection I and B for subsection II. The decision rule inferred from this choice pattern is as follows: pick the choice with the larger N for every comparison except the judgment of strength of relationship where r is the statistic presented, in which case choose the result with high r, or possible with the small N. An examination of the MVS's for this group confirms this main rule, except for the possible use of small N in Problem 6. It further suggests the use of low p as a secondary factor in Problems 3 through 5.

Table 42

CHOICE RESPONSES FOR PG 8*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
49	1	0	1	0	1	0	0	1
52	1	0	1	0	1	0	0	1

*A = 0; B = 1

PG 9 has three subjects. Data for this group is presented in Table 43. Subjects in this group chose B for all personal assurance comparisons. For strength of relationship comparisons they followed different patterns depending on whether t or r was involved. For those that presented t, they chose A for both subsections. For those that presented r, they chose A for subsection I and B for subsection II.

Table 43

CHOICE RESPONSES FOR PG 9*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
6	1	1	0	0	1	1	0	1
48	1	1	0	0	1	1	0	1
79	1	1	0	0	1	1	0	1

*A = 0; B = 1

The decision rule inferred from this is: favor low p for all judgements of assurance, with large t or small r possible adjuncts to p; favor low t for strength of relationship judgments with t (actually high p was also possible, but was not inferred as a decision rule because it was seen as a very unlikely basis for any preference); and favor high r or perhaps small N (or both) for strength of relationship with r. The MVS's for this group provided some confirmation but did not show the clear-cut confirmation seen in most of the other PG's. An inspection of the actual directional preference scores for each of the three constituents of this group indicated striking inconsistencies from subsection and certain oddities in responses (e.g., favoring a high p in every other subsection). This suggests that directional preference scores for this

group were not reliable and that these subjects show some confusion in their responses to this part of each problem.

PG 10 also had only one subject. Data for this subject is presented in Table 44. This subject followed the same pattern for all comparisons involving r, regardless of judgment type. A was chosen for subsection I's and B for subsection II's. For comparisons involving t the subject chose B for the two assurance comparisons and A for the two strength of relationship comparisons. From this the following rule was inferred: use high r (possibly small N, too) for all comparisons with r; use low p or high t for assurance comparisons with t and low t for strength of relationship comparisons with t. Thus, the inferred rule showed a stable preference for high r regardless of judgment, but a reversal in preference for different judgment types when t was presented instead. The MVS's clearly confirm the preference for high r (though not for small N as an adjunct) and show inconsistent support for the rule inferred in problems with t.

Table 44

CHOICE RESPONSES FOR PG 10*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
2	1	1	0	0	0	1	0	1

*A = 0; B = 1

PG 11 is presented in Table 45. The three subjects in this group chose B for all assurance comparisons. For strength of relationship judgments they chose A for subsection I's and B for subsection II's.

Table 45

CHOICE RESPONSES FOR PG 11*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
8	1	1	0	1	1	1	0	1
24	1	1	0	1	1	1	0	1
29	1	1	0	1	1	1	0	1

*A = 0; B = 1

The decision rule inferred from this is: pick the choice with low p for all assurance comparisons; for strength of relationship comparisons pick the choice with small N for those with t, and pick high r for those with r. MVS's clearly confirm this. They also suggest the use of low p as a secondary factor to N in strength of relationship with t.

The data for PG 12 is presented in Table 46. The one subject in this group chose B for all strength of relationship comparisons. For assurance comparisons, A was chosen for those with t and for those with r, A was chosen for subsection I while B was chosen for subsection II. The rule inferred from this is: favor low t for assurance with t, and high r for assurance with r; favor low p for strength of relationship for both t and r comparisons. This was supported by MVS's, though not completely confirmed.

Table 46

CHOICE RESPONSES FOR PG 12*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
67	0	0	1	1	0	1	1	1

*A = 0; B = 1

The data for the three subjects in PG 13 is presented in Table 47. With the exception of subject #1 this group chose B for all comparisons with r. For comparisons with t they chose B for subsection I's and A for subsection II's, again making no distinction between judgment types.

Table 47

CHOICE RESPONSES FOR PG 13*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
3	1	0	1	0	1	1	1	1
17	1	0	1	0	1	1	1	1
1	1	0	1	0	0	1	1	1

*A = 0; B = 1

From this it was inferred that subjects favored large N for all comparisons with t and low p for all comparisons with r. MVS's clearly confirmed the rule for t but not for r. A look at directional preference scores for all subjects on r comparisons suggested that subjects used somewhat inconsistent combinations of N, r and p for these four comparisons with each subject using a somewhat different combinatorial approach. This is one of two PG's for which part of the inferred choice rule is disconfirmed. Also, it is the only group where subjects appeared to use a combinatorial approach, weighing all three factors together into their decision. By and large, subjects used one primary overriding factor at a time in their choices in Problem 3 through 6.

PG 14 has two subjects. Data for this group is presented in Table 48. Subjects in this group chose B for all comparisons except subsection II of Problem 3. From this it was inferred that subjects

used small p for all comparisons except Problem 3 (assurance judgments with t), and large t is suggested as a possible adjunct to p. For Problem 3 subjects appeared to use large N. MVS's confirmed the use of low p for Problems 4 through 6 and disconfirmed the possible use of large t as an adjunct. They also partially disconfirmed the use of large N for Problem 3, where subjects showed inconsistency between subsections, using low p for subsection I, and large N for subsection II. Thus, with one inconsistency subjects in this group used low p as their choice decision rule.

Table 48

CHOICE RESPONSES FOR PG 14*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
9	1	0	1	1	1	1	1	1
14	1	0	1	1	1	1	1	1

*A = 0; B = 1

PG 15 has one subject. Data is presented in Table 49. This subject chose B for all comparisons with t as well as for subsection I's with r. For subsection II's with r, the subject chose A. From

Table 49

CHOICE RESPONSES FOR PG 15*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
32	1	1	1	1	1	0	1	1

*A = 0; B = 1

this it was inferred that low p was used as a decision rule in all t's, and large N was used in all r's. This was confirmed by the directional preference scores (the MVS's) for this subject.

PG 16 contained those nine subjects who with two exceptions chose B for all comparisons. From these choices it was inferred that subjects used low p as a rule for all comparisons. This was quite clearly confirmed by the MVS's. Data for this group is presented in Table 50.

Table 50

CHOICE RESPONSES FOR PG 16*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
25	1	1	1	1	1	1	1	1
26	1	1	1	1	1	1	1	1
39	1	1	1	1	1	1	1	1
62	1	1	1	1	1	1	1	1
68	1	1	1	1	1	1	1	1
71	1	1	1	1	1	1	1	1
76	1	1	1	1	1	1	1	1
47	1	1	1	0	1	1	1	1
28	1	1	1	1	1	0	1	1

*A = 0; B = 1

PG 17 was the residual group of three subjects that could not be suitably classified into any other PG. These were subjects with odd or unclear decision rules, where inconsistency across problems and subsections was such that it made little sense to put each into its own PG. Data for this group is presented in Table 51.

Table 51

CHOICE RESPONSES FOR PG 17*

Subject Number	Problem Number							
	3I	3II	4I	4II	5I	5II	6I	6II
4	1	1	0	1	1	0	0	1
20	1	0	0	0	1	0	0	1
55	1	0	1	0	1	1	0	1

*A = 0; B = 1