NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

The Influence of Undergraduate Research Methods and Statistics Courses on the Transfer of Reasoning Skills to Everyday Events, and Belief in the Paranormal

Davina Mill

A Thesis
in
The Department
of
Psychology

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Arts at
Concordia University
Montréal, Québec, Canada

December 1990

ISBN 0-315-68707-X

ABSTRACT

The Influence of Undergraduate Research Methods and Statistics Courses on the Transfer of Reasoning Skills to Everyday Events and Belief in the Paranormal

Davina Mill

The present study examined the effect of academic training in scientific methodological and statistical principles on reasoning about real-life-type issues. The hypothesis that this training would reduce belief in the paranormal was also tested. Seventy-six Concordia University psychology or applied social science undergraduates enrolled in introductory research methods (RM) and statistics (STAT) courses were randomly assigned to one of three conditions. A baseline group was tested during the second week of classes; another group was assessed after having completed RM and STAT courses; a third group, in addition to completing these courses, received three 20 min. tutorial sessions which emphasized the applicability of the principles they learned in the courses to everyday issues. A further 19 humanities undergraduates were included as a comparison group of students in a non-scientific academic discipline. Measures on reasoning about everyday-type events and belief in the paranormal were obtained. Findings indicate that RM and STAT courses, unless coupled with additional training, do not significantly enhance students' reasoning skills about everyday issues or reduce their willingness to endorse belief in paranormal phenomena, as compared to baseline levels. The results are not encouraging to those who view improvement of general reasoning skills as a fundamental goal of such courses, and suggestions for improvements in the curriculum are offered. Furthermore, students from the humanities discipline performed at levels similar to that of the psychology / applied social science subjects in the baseline condition, and showed significantly higher levels of belief in the paranormal as compared to all the other groups.

Acknowledgements

I am indebted to my thesis supervisor, Dr. Tom Gray, not just for his terrific contributions on my thesis, but also for all the encouragement, guidance, and support he has given me throughout my years at Concordia. Tom, your style and humor have helped to make my graduate experience as painless as possible, and actually quite a bit of fun.

I would like to thank Dr. Jane Stewart for the time she put into being a member of my thesis committee, and to Dr. Syd Miller, whose helpful suggestions really jazzed up my thesis – even if it did mean entering the wacky, wonderful world of multivariate statistics. I'm sure I'm a better person for it.

I would also like to extend thanks to the teachers of the research methods courses from the 1988 Fall semester, who permitted those of us involved in the study to invade their classrooms in pursuit of subjects. By the way, David, Pauline and Christine, as you all know, none of this would have been possible without you guys (I have yet to erase the blackboard!).

I'd also like to thank Laura Schleiffer, aside from being a terrific friend, generously shared with me her knowledge in correct grammatical structure and how to survive graduate work (♫"Oh Lord, won't you give me a Master's degree..."♫). Plus a big thanks to all my friends who have stood by me all these years.

Aunty Esther, your nimble and willing fingers were, once again, greatly appreciated.

And of course, acknowledgement goes out to my incredible parents, Dawn and Marvin Mill, who have shown me never-ending support in all that I do, and whom I love with all my heart.

I'd like to dedicate this thesis to my Bubby, whose strength and determination have been an inspiration to us all.

Table of Contents

List of Figures

APPENDICES

The ability to distinguish badly documented claims from well supported propositions, namely, to recognize the difference between propaganda and accurate information, is of great importance in coping in today's world. The skeptical attitudes needed to counteract human gullibility, however, are often not found at a very sophisticated level, even in well-educated people. In general, people are poorly prepared for evaluating the quality of the information they receive.

Research has shown that the strategies people employ in solving many types of problems often do not respect required statistical principles. For example, the work of Kahneman and Tversky (1973; Kahneman, Slovic, & Tversky, 1982; Tversky & Kahneman, 1974) has demonstrated that people often violate the law of large numbers (LLN; i.e., that sample values of parameters approach population values as a function of the size of the sample), neglect base rate information (i.e., background data against which the probability of an event should be judged), and ignore the regression to the mean principle (i.e., the tendency of extreme scores to move towards the mean as a function of repeated testing), when solving a variety of problems. The general difficulty people have with assessing contingency relationships has also been well documented (Allan & Jenkins, 1980; Alloy & Tabachnik, 1984; Jenkins & Ward, 1965; Smedslund,1963).

In a succession of studies, Kahneman and Tversky (1972, 1973; Tversky & Kahneman, 1983) have shown that much inductive reasoning frequently relies on intuitive, nonstatistical judgmental heuristics which help to reduce complex inferential tasks to simpler judgmental operations. Though these heuristics often lead to correct conclusions, they sometimes run counter to statistical and/or logical principles and lead people into erroneous judgments.

One such heuristic, the representativeness heuristic, allows the individual to reduce many inferential tasks to what are basically simple similarity judgments (Nisbett & Ross, 1980). Thus, one event may be said to be representative of another to the degree that the first event resembles the second. This heuristic involves relying on one's prototypes, stereotypes or past experiences in generating potential solutions to a problem; one compares

alternative options and selects the one that best "fits" one's representation of the current problem.

While the representativeness heuristic is often useful, problems arise when people use it as the exclusive basis for making an inference. Other necessary information (e.g. statistical) is often ignored. For example, Tversky and Kahneman (1973) tested the hypothesis that people's use of the representativeness heuristic would lead them to violate the conjunction rule. This rule states that the probablity of a conjunction cannot be greater than the sum of the probabilities of that conjunction's constituents (i.e., $p[A\&B] \leq p[A] + p[B]$). In one experiment, students were given the following description:

> Linda is 31, single, outspoken, and very bright. She majored in philosophy in college. As a student, she was deeply concerned with discrimination and other social issues, and participated in anti-nuclear demonstrations.

Students were then asked which statement was more likely; (a) Linda is a bank teller, or (b) Linda is a bank teller and active in the feminist movement. Eighty-five percent of the respondents selected option 'b'. In doing so, the respondents failed to consider that the number of bank tellers in the overall population must be at least equal to if not greater than the number of bank tellers who are active in the feminist movement. These researchers explained this occurrence in that option 'b' better represented Linda's description than option 'a'.

Demonstrations of the misuse of another heuristic, the availability heuristic, have also been well documented (e.g., Tversky & Kahneman, 1973). This heuristic is based on the finding that when estimating the frequency or likelihood of events, people often rely on the salience or availability of the events in memory. There are many factors uncorrelated with frequency or probability, however, which can influence how an event is perceived or recalled. For example, vivid information is more likely to be retained than less vivid information (Nisbett & Ross, 1980). Furthermore, information that is congruent with one's preconceptions is also more likely to be recalled and bias the consideration of subsequent evidence (Chapman & Chapman, 1969;

Greenwald, Pratkanis, Lieppe, & Baumgardner,1986). The term selective learning has been used to describe the process in which the learning and retention of attitude relevant information is most likely to occur when that information is in agreement with prior beliefs. When, for example, people have a prior belief about the causal relatedness of two variables, they may perceive the two variables as co-occurring in sets of observations in which there is no objective correlation, or even in sets with a negative objective correlation, in which the occurrence of one variable actually predicts the absence of the other (Lord, Ross, & Lepper, 1979; Nisbett & Ross, 1980). Consequently, factors besides actual frequency, statistical probability, or true causal relations affect the memorial availability of objects and events.

Although judgmental heuristics can often lead to reasoning errors, they are functional in a very important way; they help people avoid generating innumerable fruitless hypotheses in their search for useful generalizations. Moreover, although errors in judgment and erroneous inductions are often made, people often do reason quite well. The degree to which use is made of more formal statistical considerations appears to depend, however, on a number of factors. Some researchers (Fong, Krantz, & Nisbett, 1986; Nisbett, Krantz, Jepson, & Kunda, 1983) have suggested that at least three factors operate individually and, perhaps more often together, to increase people's tendencies to apply statistical reasoning, namely (a) clarity of the sample space and the sampling process, (b) recognition of the role of chance in producing events, and (c) the cultural prescriptions to think statistically.

When the sampling space and sampling process are clear, it is easier to see what knowledge is relevant and to conceptualize the observations as constituting a sample (Nisbett et al., 1983). For example, randomizing devices (e.g., dice rolling, card shuffling) are usually designed so that the sample space for a single trial is obvious and the repeatability of trials is salient, making it easier to conceptualize the observations as constituting a sample. Sampling space for objectively measurable events, such as sports contests and various achievements (e.g., job, academic), however, are less clear, thus reducing the likelihood that these events will be approached statistically. Such events, however, are still sufficiently "codable" (often, in fact, using numbers) that

people can apply a formal rule such as the "law of large numbers" (LLN). On the other hand, for purely subjective events (e.g., judgments about a person's friendliness or honesty), it can be difficult to define what constitutes an event. The sample spaces are often unclear, and repeatability not so obvious.

Fong et al. (1986) found that changing the problem domain, and consequently the clarity of the sampling space and process, did affect the likelihood of applying statistical reasoning to the problem. In their study, subjects were asked to comment on brief scenarios in which a statistical criticism was warranted. Examples were drawn exclusively from one of the following broadly defined domains: (i) manifestly probabilistic problems (e.g.,random generating devices), (ii) objectively measurable events (e.g., achievements of some kind) or (iii) subjective judgments (e.g., assessing someone's sense of kindness). Subjects were most likely to employ statistical reasoning for probabilistic problems (75%), less likely to do so for objective problems (48%), and least likely for subjective problems (33%). Nisbett et al. (1983) found that by manipulating the codability of events in such a way as to make the events more easy to compare by suggesting a unit of comparison, led to an improvement in people's ability to make use of more formal statistical considerations.

People's ability to make appropriate generalizations also depends on how familiar they are with the sampling space (i.e., the extent to which they believe the population the sample comes from is homogeneous). For example, Quattrone and Jones (1980; cited in Nisbett et al.,1983) demonstrated that people are more likely to make generalizations about members from groups with which they are less familiar (as opposed to those with which they are more familiar), as they do not recognize the group's general variability. In addition, when people are instructed to contemplate the central tendencies of the population under consideration (i.e., to consider the degree of variability that exists in the group), changes in the willingness to generalize were noted (Nisbett et al., 1983).

The recognition that chance factors might be operating is also important in encouraging the application of statistical analysis to a problem. For example, people who realize there is a random element in the way a football bounces

may be more likely to recognize the role of chance factors in the outcome of a game. In many social situations, however, it may not be easy to recognize the role played by chance factors resulting in less use of statistical reasoning to these problems. It should follow that expertise in a given field is likely to bring about a greater recognition of the chance factors at work in that domain.

The third, more general factor, noted by Fong et al. (1986), that may influence the use of a statistical approach about a given event is concerned with the extent to which the culture encourages the use of statistical reasoning about the event. For instance, strong religious beliefs in certain communities might discourage reasoning about some events and instead encourage belief in the possibility of supernatural intervention. For example, influences in medieval European culture reinforced the persecution of "witches", and today's resurgence of Christian Fundamentalism has led many to not question various religious beliefs (e.g., the belief in Creationism), even in the face of contradictory evidence (e.g., the theory of evolution); (Alcock, 1981).

Formal Discipline: Transfer versus Domain Specificity

One question that has caught the attention of many throughout history is concerned with whether people can be taught to reason more effectively. Plato stated the concept of formal discipline, which basically holds that "reasoning can be improved as a result of teaching rules within a particular field, which then generalize outside the bounds of that field" (Lehman & Nisbett, in press). Later, Roman philosophers added the study of grammar and the training of the faculty of memory to the list of exercises that were useful for formal discipline. The medieval scholars then added an emphasis on logic (e.g., the syllogism), and the humanists later added the study of Latin and Greek. The formal discipline view ultimately became so immoderate that educators were able to advocate the teaching of a content field strictly for its discipline or exercise properties. As one mid-19th century educator stated, "The acquisition of a language is educationally of no importance; what is important is the process of acquiring...The one great merit of Latin as a teaching instrument is its tremendous difficulty" (cited in Nisbett, Fong, Lehman, & Cheng, 1987).

For much of the 20th century two different positions argued against the "formal discipline" view. The first held that reasoning skills were highly domain-specific and nongeneralizable (Thorndike, 1906, cited in Nisbett et al., 1987). Thus, adherents to this view did not deny that people could be taught reasoning skills, but claimed that what would be learned would be generally applied to other situations only if they were very similar. Thus, it was asserted that all reasoning takes place by domain-specific rules and that more abstract rules play no part in everyday reasoning. The second view, while agreeing with the formal discipline model that reasoning skills are generalizable and based on abstract rules, assumed that the process of acquiring these skills could not be accelerated through training; instead the development of the skills depended on the normal course of cognitive maturation (Inhelder & Piaget, 1958).

Although the two anti-"formal-discipline" views differed considerably, both were in agreement that teaching abstract rules would be, for the most part, ineffective and that training in a given domain should produce little transfer to other domains, beyond what would have naturally occured.

Current research, however, seems to show that this pessimistic view of the trainability of inferential rules is mistaken. For example, Fong et al. (1986) examined the effects of brief training sessions in the "law of large numbers" concepts (LLN) both in the abstract and with examples drawn from a given broad domain. Abstract instruction in the LLN consisted of defining the notions of sample, population, and parameter, and illustrating that as sample size increases, the sample usually resembles the population more closely. Training on examples consisted of showing how to use the LLN to solve a number of problems from broadly defined domains. Some subjects received no training at all, some received abstract rule training only, some examples training only, and some received both rule and examples training. Each of the two separate training procedures took less than 30 minutes. All subjects were then tested on their ability to apply the LLN to a variety of problems.

The results of this study indicate that even very brief training in the LLN can increase both the quality and frequency of statistical reasoning as applied to everyday problems. Furthermore, subjects who received both rule and examples training were more likely to use statistical reasoning in their answers than those who received only rule training or examples training alone. Those

receiving no training performed the poorest. These results support the view that people can appreciate the LLN in the abstract sense and that training can facilitate transfer to a wide range of problem content. Different results, however, have been reported in assessing the effectiveness of deductive reasoning training. Cheng and Holyoak (1985) and Cheng, Holyoak, Nisbett, & Oliver (1986) have found that teaching deductive reasoning skills, such as the proper use of the material conditional (e.g., if p, then q), through either formal rule-based or example-based approaches is ineffective unless both approaches are coupled. These researchers conclude that the disparity in results in teaching the LLN versus the conditional is due to the fact that people possess a formalized version of the LLN in their problem solving repertoire, whereas they do not have one for the material conditional.

What, then, is the basis for acquiring one group of concepts (e.g., the LLN) more readily than another (e.g., the material conditional)? Cheng et al. (1986) have proposed that differences in concept attainment reflect a concept's range of applicability and, hence, its pragmatic value. A statistical concept like the LLN is meaningful at a context-independent (i.e., abstract) level, whereas the usefulness of a deductive concept such as the material conditional is much less obvious at a purely abstract level.

While it appears that people do not possess an inherent appreciation for the conditional in the same way that they do for the LLN, some researchers (e.g., Cheng et al., 1986; Lehman et al. 1988; Nisbett et al., 1987) have suggested that people do, in fact, possess deductive reasoning skills, but that these skills are more schematic in nature. That is, these reasoning skills, referred to as **pragmatic inferential rules** (PIRs) are abstract in that they are not tied to any content domain, yet they are defined in terms of classes of goals and types of problems familiar to those encountered in everyday-type situations. For example, the standard conditional syllogism could readily be represented by a PIR called a permission schema as follows: If you wish to do p, then you must first do q (e.g., if you wish to drink, then you must be over 18 years). As it happens, the procedures for checking whether an infraction of a permission has occurred are the same as those for checking whether the conditional rule has been violated. For example, one must establish that q has occurred (permission

has been obtained) when one finds that action p has been carried out, and one must establish that action p has not been carried out when one knows that q (permission) has not occurred. Thus, when people reason in accordance with the rules of formal logic, they normally do so by using PIRs that happen to map onto the solutions provided by logic (Nisbett et al., 1987).

Cheng et al., (1986) have found that training in the material conditional is effective only when abstract principles are coupled with PIRs, which serve to elucidate specifically the mapping between abstract principles and concrete instances. Training in rules of logic without such examples fail to significantly improve performance. Even training in an entire one-semester course on logic had little effect on improving reasoning on the conditional (Nisbett et al., 1988). This is consistent with the view that the material conditional is not part of people's intuitive reasoning repertoire, and hence they lack the ability to put abstract rule training to use. Training on PIRs seems promising as an approach to improving everyday reasoning.

The results of the laboratory findings outlined thus far are optimistic insofar as instruction in abstract rule systems can improve one's reasoning on concrete examples, though training which presents the rule and also accompanying examples is more effective than either training method alone. Furthermore, learning some principles (e.g., the material conditional) requires accompanying training in PIRs.

As schooling traditionally aims to prepare students for life beyond academia, it is often considered the appropriate forum in which to foster people's reasoning skills and hone their critical abilities. Two recent influential studies on the state of American higher education have stressed the importance of fostering students' ability to think critically as one of the indispensible products of an undergraduate education (cited in Pascarella, 1989). Though inferential rules and the transfer of these rules can be taught in the laboratory, one can still ask if higher education does effectively promote the achievement of these basic reasoning skills, especially in applying them to everyday contexts.

Perkins (1985) examined the extent to which postprimary education enhances informal reasoning skills by comparing performance levels of first and fourth-year high-school students, college students, and graduate students on

the quality of their oral arguments concerning various everyday issues. As expected, graduate students in general scored higher than college students, who scored higher than high school students. Of course, these findings were confounded with selective admission procedures; in support of this idea, a regression analysis revealed that IQ was the most influential variable in predicting reasoning performance. Years of education was only a modestly significant predictor of ability. Perkins (1985) suggests that higher education fails to significantly improve general reasoning ability because it primarily focuses on context-specific knowledge rather than on teaching students to apply divergent or creative thinking outside of their area of expertise. This conclusion is not necessarily inconsistent with the theory of formal discipline; rather it suggests that if one of the aims of the educational system is to teach students how to reason better, then changes in the system are needed.

Other researchers (Fong et al., 1986) have found that as years of schooling increase, application of statistical reasoning to everyday events improves, both in quality of the answer given and in frequency. These researchers examined the effects of differing amounts of statistical education on answers to a given problem. Subjects who had no background in statistics almost invariably responded with nonstatistical answers. Subjects who had taken one statistics course gave answers that included statistical considerations about 20% of the time. Beginning graduate students in psychology, who had taken one to three courses in statistics, gave statistical answers about 40% of the time. Doctoral-level scientists at a research institution gave statistical answers about 80% of the time. Training affected the quality of statistical answers as much as it did their frequency.

The results just described, however, were based on correlations, and again can easily be confounded with selective admission procedures. To better understand the effects of higher education in an experimental context, the same researchers (Fong et al.,1986) conducted a telephone survey of opinions about sports. The subjects were males who were enrolled in an introductory statistics course. Some subjects were randomly selected and "surveyed" during the first two weeks of the term, the others at the end of the term. Subjects were asked questions for which a statistical answer would be very appropriate. The statistics

course markedly increased the frequency and quality of statistical answers to these questions.

Results from another study (Leshowitz, 1989) found similar results. First, it was noted that few first year university students in psychology and philosophy courses demonstrated even a semblance of acceptable methodological reasoning about everyday life events as described in newspaper and magazine articles. Reasoning refered to the "rules about assessing causality, rules for generalizing, rules for determing argument validity, and rules for assessing the probativeness of evidence" (p.441). A five-week introductory experimental psychology course in which a large segment was dedicated to teaching the rules of inductive reasoning with intensive application to everyday life events, however, resulted in a dramatic improvement in performance on the same test. Students from an introductory philosophy course did not show significant improvements. These results lend further support to the view that the application of principles of statistical-methodological reasoning to everyday situations can in fact be taught to undergraduate psychology students when given the proper training.

Studies comparing the effectiveness of different graduate programs on everyday reasoning have also demonstrated that certain types of training can lead to significant improvements in reasoning. For example, Lehman, Lempert, and Nisbett (1988) studied the effects of two years of graduate education in four academic fields: psychology, medicine, law, and chemistry. Graduate students in these fields were tested on several different kinds of inferential skills: (i) statistical reasoning about both scientific content (i.e., statistically flawed studies in the natural and social sciences) and everyday life content (i.e., sports events), (ii) methodological reasoning dealing with different types of confounded variable problems, for example, self-selection problems, sample bias problems, and inferential uses of control groups, for both scientific content and real life content, and (iii) ability to solve both arbitrary and meaningful problems involving the conditional and biconditional. Two different studies were conducted, one with a cross-sectional design (i.e., first-year students in each field were compared with third-year students) and one with a longitudinal design (i.e., first-year students in each field were tested and, after two years of training, tested again).

Initial differences among the three groups were very slight for all types of reasoning studied. The effects of training on ability to use statistical rules and confounded variable rules, however, were marked for both psychology and medical students, both for scientific problems and for everyday life problems. Neither law students nor chemistry students improved in reasoning using either statistical or methodological rules, for either type of content, in either cross-sectional or longitudinal designs.

These researchers also found that law, medical, and psychology, but not chemistry graduate students' performance on questions requiring conditional reasoning improved from first to third year within their program. Though none of these disciplines provide training in formal logic, law, medicine and psychology all provide training in PIRs which are analogous to the material conditional. For example, the statement that factor A causes factor B but that factor B does not cause factor A is structurally related to the conditional statement, and taught in detail in both psychology and medicine, and law students are taught about contractual relations which resemble permission schemas. On the basis of these findings, these researchers concluded that abstract statistics rules that are taught in the probabalistic sciences, but not in the mainly deterministic sciences such as chemistry, are generalized outside the bounds of their immediate domain and into the realm of everyday life events, and that training in PIRs improves reasoning on the conditional. Thus, the different rule systems emphasized in the different fields result in distinct patterns of inferential gains.

Other studies have corroborated these findings. Gray and Mill (1990), for example, asked 96 Graduate students in biology or English to read one of three abstracts that made a particular claim. For example, one of the abstracts concerned decreases in the incidence of dental caries in a population that had fluoride in the water. The abstracts differed in their apparent scientific relevance, but an important feature of all of them was that they did not contain crucial, comparative or "control-group-type" information. A measure of "critical abilities" assessed how readily the respondents recognized that crucial information was missing. Although biology students in general required fewer cues, they did not perform significantly differently from the English students on the less scientific texts. The students were also asked to read the following passage: "Suppose

you had heard a report that researchers had come up with a new drug and that 75 of the 100 patients given the drug got better". Students were then asked to comment on the effectiveness of the drug. Approximately 50% of the biology students, and only 25% of the English students, managed to detect that the fundamental comparative information was missing in the drug effectiveness task. It was suggested that the type of training received had an impact on critical thinking as measured by those tests; English students are not typically expected to solve problems requiring methodological reasoning, whereas biology students often deal with problems such as controlling design flaws in their discipline.

Finally, a comparative group of psychology graduate students was examined on the same tests (Gray, in press). Considering that training in methodological and statistical reasoning skills is a trademark feature of the social sciences, it was disquieting to find that they performed only marginally better than the biology students. These findings were taken to indicate that the unimpressive performance in everyday-type reasoning is due to the context-specific training students generally receive in their respective fields. That is, training typically does not emphasize the general applicability of the rules taught in the courses to other domains. As such, if part of the problem lies with how the courses are taught in the school system, then improving educational practices should maximize students' ability to apply these problems to their everyday experience.

Belief in Paranormal Phenomena

Research findings indicate that high percentages of people believe in phenomena for which there is no evidence that would generally meet scientifically acceptable criteria of trustworthiness (Alcock, 1981; Gray, 1985; Harrold & Eve, 1987). Belief in extrasensory perception (ESP), for example, has been found to be moderate to strong in 80-90% of the general population (Gallup, 1978). These high levels of belief are found even within university educated populations. For example, Gray and Mill (1990) found that 80% of their graduate sample believed in at least one paranormal phenomenon. Levels

of belief were even higher for undergraduates (Gray, 1985; Gray, 1990a,b).

What factors contribute to these high levels of belief in phenomena for which there is no scientific evidence that meet trustworthy standards in support of these phenomena? It is likely that there are many factors that contribute to these beliefs (see e.g., Alcock, 1981; Singer & Benassi, 1981). One explanation that has been proposed to account for belief in these phenomena is that they emerge as a result of poor abilities in assessing evidence. To illustrate this point, many people, for instance, have encountered a situation in which they think about an event (e.g., thinking about a particular song) and soon after actually experience it (e.g., suddenly the song is heard on the radio). For some people, this event is compelling, and they are inclined to ascribe to it a paranormal cause (e.g., i was thinking of song X, and lo and behold, it came on the radio; I must have used telepathy to inform the disc jockey of what to play). These people, however, have ignored the far greater number of non-occurrences (e.g., considerations of all the times one had thought of a song and it was not played on the radio). Failure to consider these negative instances can be understood in context of people's general difficulty in assessing the base rate of an event. For example, believers do not think that the song situation is perhaps 1 in a 1000 chance occurrence. Instead, it is taken to represent a very interesting situation in and of itself. Or else, they selectively attend to information consistent with their original hypothesis, and ignore disconfirming evidence. Findings from a series of studies (Jones & Russell, 1980; Russell & Jones, 1980) have supported this view. An asymmetry in the selective attending of believers and disbelievers in paranormal phenomena was demonstrated, in that believers' ratings of their own ESP abilities on a card sorting task, unlike those of the skeptics, were unrelated to feedback received on their actual performance on the task.

As an extension of the findings from the literature on human reasoning, some researchers have begun to investigate the link between reasoning skills and belief in the paranormal, as providing at least a partial explanation for high levels of belief in these unsubstantiated phenomena. Some investigators have examined the relationship between levels of belief in the paranormal and performance on various reasoning and critical thinking tasks, and have found

moderate, yet statistically significant correlations (Alcock & Otis, 1980; Gray & Mill, 1990; Wierzbicki, 1984).

Gray (1985) has suggested that one of the reasons why people may believe in various paranormal events is that they are unaware of what constitutes "good" evidence. He further suggests that they do not automatically invoke what he calls the "control-group-way-of-thinking" about evidence. This way of thinking is basically the tendency to systematically evaluate claims using comparative information. Applying this type of evaluation to the evidence whenever possible, Gray suggests, should lead to less errors caused by ignoring statistical considerations (e.g., base rates or sample size) and should help to increase awareness of the possibility that other plausible alternative hypotheses may account for the results (e.g., sampling bias, confounded variables, etc.).

In order to test the hypothesis that these beliefs are at least partly a result of respondents' ignorance of alternative, non-paranormal explanations, and their ignorance of what constitutes "good" evidence, Gray (1985) offered undergraduate students a one semester course dealing with critical examinations of the quality of evidence supporting paranormal phenomena. He measured students' beliefs prior to the course, after the course, and at a one-year follow-up period. Attempts to reduce these beliefs produced moderate, though statistically significant and to some extent durable, reductions in undergraduates' willingness to endorse belief in various paranormal phenomena. Though these decreases were observed, Gray noted that the percentage of believers was still quite high (e.g., approximately 85% believed in ESP prior to the course as compared to 55% following it). As many students simply did not change their beliefs, it appears that belief in the paranormal is not only a result of ignorance as to what constitutes good evidence; these beliefs draw their support from more than one source. As beliefs are probably formed over many years and due to many factors, the moderate reductions in belief after just one course perhaps may be considered as relatively promising.

Though levels of belief are typically high for many paranormal phenomena, there is some evidence to suggest that the stronger the background in science (where the principles of statistical and methodological reasoning are emphasized), the lower the levels of belief in the paranormal. In

one study, Gray (1990a) reported lower levels of belief for undergraduate students in the natural sciences as compared to those students in English or psychology. The differences were not very large, but perhaps this could be attributed to the fact that the advantages of academic training in a specific field do not fully manifest themselves until graduate and post-graduate levels of training. In another investigation, Gray and Mill (1990) found that graduate students studying English had significantly higher levels of belief in paranormal items than did biology graduate students. Furthermore, Otis and Alcock (1982) found that English professors were the only group out of eight university departments to be consistently high in acceptance of belief in four extraordinary belief categories (e.g., spirits, fortune telling, psychic healing, and religion).

These findings lend some support to the notion that paranormal beliefs may partly result from ignorance of the methods and skills useful in evidence collection and evaluation, considering the lower levels of belief found for the science students. Of course, one problem inherent in such comparisons is that self-selection effects may be present. That is, perhaps people with greater skeptical tendencies are drawn to the field of science, rather than the science background producing more skeptical thinkers. Unfortunately, these studies were not designed to answer these questions, and no longitudinal studies going back to students' earlier scholastic training (i.e., before they have selected their field of study) are available.

It should still be emphasized that, as mentioned earlier, the science graduate students who have supposedly received a lot of training that specifically emphasizes the scientific method, are still not very skeptical of these phenomena, as evidenced by the high numbers of those who still believe in these phenomena. Therefore, it appears that these students do not approach evidence concerning the paranormal with the same critical stance that they would apply to evidence in their own field of study.

The aims of the present study:

One objective of the undergraduate introductory psychology research methods (RM) and statistics (STAT) courses, aside from teaching the students

how to apply this knowledge to psychological research, is to assist students in honing their reasoning and critical thinking skills at a general level. Reasoning and critical thinking skills in this case are defined as the ability to apply various statistical and methodological principles to problems encountered. The present study was designed primarily to investigate whether or not the knowledge gleaned from these courses would transfer readily to other domains (i.e., to everyday-type events). Based on the findings in the literature (e.g., Fong et al., 1986; Gray & Mill, 1990; Leshowitz,1989) and an informal examination of these courses, it was expected that these courses would not have a very large positive impact on the transfer of reasoning skills to other domains. In addition, the effect of brief tutorial sessions on reasoning ability was also examined. These sessions were developed based on the principles presented in the literature (e.g., Cheng et al.,1986; Nisbett et al., 1983). It was expected that the additional tutorials would be necessary to improve reasoning skills above baseline levels (i.e., levels observed before taking the courses).

Another objective was to assess the effect of the different types of training on levels of belief in five paranormal phenomena. It was hypothesized that enhancing reasoning skills would lead to lower levels of paranormal belief, and that reasoning training specifically using paranormal examples would lead to the greatest reduction in levels of belief. An additional aim was to assess the relationship between statistical and methodological reasoning skills and belief in the paranormal. It was expected that reasoning ability would be inversely related to belief in the paranormal.

A comparison group comprised of humanities undergraduates who had not received any formal statistical or methodological training was also included, in order to assess the influence of differential academic training on both reasoning skills and belief in the paranormal. It was expected that this group would show the lowest levels of reasoning ability and the highest levels of belief in the paranormal.

Method

Subjects

Seventy-six (17 male, 59 female) Concordia undergraduates enrolled concurrently in introductory Research Methods (RM) and Statistics (STAT) were recruited in the first week of the semester. Most of the students were registered as psychology majors but some were registered in the applied social science (APSS) program which also requires completion of these courses. A further 19 undergraduates (6 male and 13 female) enrolled in Concordia's Liberal Arts College were recruited. Most of these students were registered in a humanities major (usually english or history) and were excluded from participation if they were taking or had taken a statistics or research methods course.

The syllabuses for the RM and STAT courses were standardized across all sections of the courses.

Psychology subjects were randomly assigned to one of five experimental groups. With the permission of each professor, this was accomplished during class time by giving each eligible subject one of five different colored papers (yellow, white, blue, green, or pink) on which there was a brief description of the study (Appendix I). The color of the paper determined to which group the participant would be assigned. The humanities students were recruited from their classes too, and consequently made up a sixth group with which to compare to the psychology/APSS students' performance.

All subjects had the chance of winning a $50 or $100 prize. The general interest and pedagogical usefulness of the experience was emphasized, and subjects were reassured that individual results would be kept confidential. It was stressed that participation was voluntary and would have no bearing on their grade, and that they could discontinue participation at any time. Approved human subject protocol procedures were used throughout the study. Subjects' data were excluded from the study if they officially discontinued in the courses or if they failed the courses because they unofficially dropped out, or did not demonstrate an adequate ability to express themselves in English, a decision which was left to the discretion of the examiners. Drop out rate was very low; of all the eligible psychology/ APSS participants receiving a colored paper, only

14 people chose not to partake in the study. Attrition rates did not differ significantly across groups ($p > .05$).

Ages ranged from 18 to 44 (mean age = 22 yr).

Materials

Personal Data Form. This form consisted of 8 questions, concerning participants' sex, age, english fluency, and educational background (Appendix II).

Gray Belief Survey (GBS). This questionnaire was used to assess subjects' belief in 10 phenomena (Appendix III). Three of the items were scientifically supported topics (namely, the link between smoking and cancer, the theory of evolution, and that germs contribute to disease); two of the items were scientifically controversial (Vitamin C prevents colds, apes possess sign language); and five items consisted of paranormal phenomena (ESP, UFOs, astrology, psychic healing, and reincarnation). The first page of the survey provided subjects with a brief explanation of each item. A brief description of each phenomenon was provided with the single-page questionnaire that allowed the respondents to indicate their belief or disbelief for each item. The strength of belief was indicated by checking +1, +2, +3, or +4 corresponding to a Weak (+1) through Strong (+4) "believe" response. Similarly, respondents could check -1, -2, -3, or -4 corresponding to a Weak (-1) through Strong (-4) "don't believe" response. The phenomena were arranged (randomly) one above the other down the centre of the page. The four "don't believe" boxes were to the left, and the four "believe" boxes to the right.

The instructions stressed the importance of responding according to whether or not respondents believed in the reality of the phenomena, not just the theoretical possibility.

This questionnaire took only about 5 minutes to complete.

Wonderlic Personnel Test - Form A (1983) (Wonderlic). This test was used to obtain an estimate of subjects' general intellectual aptitude. It consisted of 50 questions of increasing difficulty, providing a broad range of problem

types (e.g., analogies, arithmetic problems, similarities, spatial relations, etc.). Completion time was 12 min. The Wonderlic correlates .91 to .93 with the Weschler Adult Intelligence Scale - Full Scale IQ (Dodrill, 1981). The test-retest reliability ranged from .82 to .94 (Wonderlic, 1983), and longitudinal reliability was measured at .94 (Dodrill, 1983).

Gray Critical Ability Test (GCAT). This test was designed to assess the subjects methodological reasoning, that is, their tendency to employ what has been called a "control-group-way-of-thinking" (Gray, 1985). The test assesses the respondents' ability to recognize that the material they have been given lacks appropriate comparative or control-group type information (Gray & Mill, 1990).

The test involves reading a brief text (approximately 200 words) and then being asked questions concerning the claims made. Two texts were alternately used in order to assess the effect of the topic on the subjects' performance. In both cases, crucial comparative information necessary to properly evaluate the claims is lacking.

One text (*Fluoridation and Declining Tooth Decay*) was an edited version of a "News and Comment" article from Science, and read as follows:

> "Fluoridation consists of raising the concentration of the fluoride ion F in water supplies to about 1 part per million with the aim of reducing dental caries (tooth decay) in children.
>
> The use of fluoride additives in water supplies or the provision of fluoride tablets to children is thought to reduce the number of Decayed, Missing, and Filled teeth (DMFT).
>
> A study from Brisbane in Australia reported a 50% reduction in caries over a 23 year period as measured by the DMFT count per child after introduction of fluoride tablets.
>
> Further support for the usefulness of fluoridation comes from a study that looked at the use of fluoride in the municipal water supply. Reductions in decay of 71% to 95% were reported over a 16 year period of fluoridation."

The second text (*Adolescent Suicide and Parental Harmony*), which was constructed for this study, and was based on the format of the Fluoride text:

> "Adolescent suicide has increased dramatically in many Western countries in the last decade. Recent attempts to discover variables that might be predictive of subsequent suicide have uncovered what might be important findings.
>
> A study from Brisbane in Australia reported that in 52% of the cases of attempted suicide in adolescents the parents reported that their marriage relationship was undergoing "moderate to serious" stress.
>
> Further support for the role of parental harmony in connection with adolescent suicide attempts comes from a similar, long-term study which reported that 71% to 95% of adolescents who attempted suicide felt that the relationship between their mother and father was "frequently strained".

In the Fluoride text, the experimenter was looking for some evidence that the subjects recognized that the rates of tooth decay in people not receiving fluoridation in similar regions over a similar time period was not included in the text. Similarly for the Suicide text, recognition of the omission of rates of attempted suicide for adolescents not experiencing parental disharmony was assessed.

The GCAT was administered on an individual basis by a trained experimenter. Participants were asked to read the short article and then given an opportunity to make an open-ended comment on the abstract they had just read. They were asked: "what do you think of the claims made in the brief article you have read?" The abstract was available to them to refer to throughout the interview.

The respondents' comments were recorded in note form and were taped. The interviewer was trained to respond to any questions in a passive, non-leading fashion. It was sometimes necessary to ask respondents to elaborate on what they said, but care was taken not to explicitly cue them towards recognition of the lack of "control-group-type information". If the respondent said "there needs to be a control group", respondents were asked to further describe

what they meant. If it was clear that the respondent spontaneously recognized that crucial information was missing, a score of "zero cues needed" was assigned.

All respondents were then given a general "priming" cue that has been termed the "Smedslund cue", so-called because of the relevance to Smedslund's (1963) landmark paper. The aim of this cue, aside from further measuring methodological reasoning, was to help provide a general "set" to encourage a critical attitude. The few respondents who had spontaneously recognized the lack of crucial information in the abstract during the opportunity for open-ended comment were also given this general cue but were not given the subsequent series of 6, more specific, cues.

This general priming cue consisted of the following question:

> "Suppose you heard a report that researchers had come up with a new drug and that 75 of the 100 patients given the drug got better. What comments would you have on the effectiveness of this drug?"

Subjects' responses were recorded and scored merely as "yes" (1) or "no" (2) depending on whether or not their answer indicated that they realized nothing could be said about the effectiveness of the drug in the absence of information about recovery rates for patients not receiving the drug (or placebo control).

Following the Smedslund cue, respondents were then asked to go back to the article they had just read. Those who had not identified the problem with the text immediately were given a series of cues aimed at eliciting comments that indicated they recognized that crucial information was missing. The cues were as follows:

a) "What do you think of the evidence the author presented?"

b) "Do you think the information provided by the author is sufficient to support the claim?"

c) "Would some other, additional information have been useful?"

d) "Do you think that the data that was presented was representative

and/or reliable?"

e) "What about some comparative information, that is, would you like some information to compare with what was presented?"

f) " Do you think it would have been important to know about changes in the rate of tooth decay in populations not given the Fluoride?" or " Would it be important to know what proportion of adolescents in general report parental disharmony (i.e., those who have not attempted suicide)?"

Cuing continued until either a response was elicited indicating the subjects' awareness for the need of comparative information, or all six cues were given.

Two scores were obtained from this test. The first measure was based on when the participant realized that other factors not mentioned in the text may have acted as confounding variables (**CONF**). The second, more stringent measure, involved determining when the subject was able to indicate the need for comparative, control-group information (**CNTL**). As such, a correct answer on the CONF measure can be seen as identifying that a problem may exist, and a solution to this problem is reflected in a right answer on the CNTL measure. Accordingly, two scores were derived, one for the control-group and one for the confounding variable. The average was calculated from these two scores and used as the measure for the text condition of the GCAT (**TEXT**). Scores ranged from 0-7; a score of 0 indicated the subject needed no cuing to recognize the need for comparative information; a point was assigned for the number of cues needed. If, after the 6 cues, the subject still did not realize the need for the necessary information, a score of 7 was assigned. Evaluating at what point the subject satisfactorily answered the question turned out to be straightforward. Inter-rater reliability amongst three researchers for CNTL and CONF was respectively calculated at .85 and .83, and at .94 and .85 allowing for a 1–point disagreement in either direction. Discrepancies were reconciled on the basis of review.

It should be emphasized that it was not necessary that the response be couched in specific scientific or social science methodological jargon. The response could be in everyday language so long as it was clear that the respondent realized that the information given in the abstract was inadequate to

support the conclusion. For example, a good CONF response to "Fluoride" could be "well, maybe over all those years eating habits or brushing habits changed. Maybe people started looking after their mouths better. Maybe fluoride was not necessary."

This portion of the test session took about 15 minutes.

Aside from the face validity this test has in measuring subjects' ability to recognize missing comparative information from an article of everyday relevance, this test has already been shown to distinguish English from biology graduate students (Gray & Mill, 1990).

Lehman Reasoning Test (**LRT**). This is a modified version of Lehman, Lempert, and Nisbett's (1988) general reasoning test (see Appendix IV). The LRT is an 11 item, multiple-choice, paper-and-pencil test that, along with the GCAT, provides a measure of participants' methodological-statistical reasoning abilities. Unlike the GCAT, the LRT is not dependent on oral expression. Again, questions were designed to address everyday-type situations. Thus, subjects did not need to be familiar with technical jargon in order to understand or answer the questions. The questions could be grouped into the three areas they addressed: knowledge of statistical principles (e.g.,law of large numbers, regression towards the mean, and base rates), methodological reasoning (e.g.,recognizing confounding variables, self-selection effects, or the need for control groups), and ability to solve problems of the material conditional. Problems on the questionnaire, although embodying the same principles as those taught in the critical thinking tutorial sessions (see next section), did not correspond directly to any problems actually presented in the tutorial sessions. Possible scores ranged from 0-12. The complete version of this test has been shown to be able to distinguish graduate students from different faculties (Lehman et al.,1988)

Design and Training Conditions

The main concern of the study was to test the effect of formal academic training (e.g., the RM and STAT courses) on students' methodological and

statistical reasoning about everyday-type problems. An additional consideration was with whether completion of the courses led to a decreased willingness to endorse belief in scientifically unsubstantiated phenomena.

It was not feasible to use a completely within-subjects design, as a practice effect on the tests would have occurred, especially on the GCAT (as all subjects know the answer by the end of the test). In addition, a pre-measurement of the dependent variables may have reduced the internal validity of the experiment by influencing the participants' experience in their RM and STAT courses (e.g., by artificially cuing the subjects as to what they should be learning in the courses in order to succeed on the dependent measures). Hence, a mixed repeated measures design was used to assess the impact of formal training on reasoning and critical ability and belief in the paranormal. Four basic conditions were established, and are as follows:

*1. Baseline Control Group (**BASELINE**)*:

This group was comprised of 20 students enrolled in the RM and STAT courses. As they were tested in their second week of classes, they were considered to have had no formal training in RM and STAT, and were used as a control group to compare to the groups who had completed the coursework.

*2. RM/STAT condition (**RM/STAT**)*:

Two groups of 14 subjects each were tested at the end of the semester after completion of the RM and STAT courses. One of the two groups in this condition took part in three, 20-min. "information sessions" that dealt with material concerning student academic affairs, employment opportunities, etc. That is, the tutorial sessions for these subjects were irrelevant with respect to information about methodological and statistical reasoning. This group received the same amount of extra attention as was given to subjects in the Tutorial condition.

*3. Tutorial condition (**TUTOR**)*:

Two groups of 14 subjects each participated in three 20-min. training sessions over the course of the semester, starting in the fifth week of classes,

taking place every other week for six weeks. These sessions were a supplement to their regular course work in the RM and STAT courses. These sessions emphasized the applicability of methodological and statistical techniques to examples encountered daily, without using technical terminology. For example, the discussion of the "law of large numbers" involved relating it to everyday-type situations, such as evaluating the claims made in a T.V. commercial or in making judgments about what characterizes a person. Throughout the sessions, it was frequently stressed that these critical thinking skills can and should be applied to everyday reasoning (e.g., assessing claims made by politicians and the media, choosing between competing brands of a product, etc.). A large emphasis was placed on the need to seek out comparative, control-group type information when assessing many kinds of evidence.

The only systematic difference between the two groups was that for one group (paraTUTOR group), approximately half of the examples dealt with issues concerning how to evaluate claims about the paranormal using a critical approach (i.e., that these paranormal events could be explained more parsimoniously in terms of "natural" causes), whereas no paranormal content was included in the examples for the other group (normTUTOR group). So, for example, in discussing the importance of base rate information, it was pointed out to the paraTUTOR group that some "psychics" use such information to get a good idea what people's likely responses will be in advance, in order to appear as if they have paranormal abilities such as telepathy, whereas the normTUTOR group received an example from a newspaper article dealing with the efficacy of matchmaking firms, in which base rate information was lacking.

The normTUTOR and paraTUTOR groups, both of which received the extra tutorials in critical thinking about everyday issues, would be combined for statistical analysis on the two reasoning tests (i.e., GCAT & LRT) to create one group called **TUTOR**. The TUTOR group would be split apart again for analysis on the Gray Belief Survey (GBS), as it would be necessary to assess whether or not training including examples pertaining specifically to the paranormal would be needed in order to employ critical thinking skills in that domain.

4. *The Humanities Group (**HUM**)*:

A group of 19 students recruited from Concordia's Liberal Arts College (i.e., the humanities discipline) in either their first or second year of study, served as a comparison condition. None of these students had taken a RM or STAT course.

Three researchers were involved in giving the tutorial sessions. Preparation included: detailed transcripts of required material being drawn up for each session, instructors becoming thoroughly familiar with the material, in order that all three instructors would be giving comparable sessions. Nonetheless, the groups were further subdivided into three subgroups for two reasons: First, to keep the groups smaller, thus making it easier to engage individual attention; secondly, to be able to counterbalance trainers across the three groups and across the three sessions in order to control for effects of possible experimenter attributes on the results.

Testing Procedure: All groups were given the same tests in the same manner, and only differed in <u>when</u> they were tested. The BASELINE group was tested in the second week of the first semester, the HUM group the following semester, while all other groups were tested at the end of the first semester before final examinations. With the exceptions of the BASELINE and HUM group, experimenters were blind with regards to which group a participant belonged.

All subjects signed the consent and lottery forms first, filled out the personal data form next, answered the GBS, then received the Wonderlic. At this point, in order to examine possible sequencing effects, participants were evenly divided so that half completed the LRT first, and the other half completed the GCAT first. Each of these halves were further split so that half of each received the "Fluoride" text and the other half got the "Suicide" text. Thus, counterbalancing techniques were employed for both critical thinking tests as well as the type of text used. The duration of testing was approximately 1 hr 15 min. All subjects were asked not to discuss the study with the other participants, and were told they could find out their personal results and/or the purpose of the experiment during the following semester.

Results

The data were analyzed using the Systat software application for Macintosh computers.

As no significant differences were observed in the means for the two texts (Fluoride and Suicide) on CNTL, CONF, or SMED scores (p = .39), data from both text groups were combined for further analyses. With regards to the possible influence that the order in which one received the GCAT and LRT might have exerted, an ANOVA was performed to rule this possibility out. No order effects were found across all measures associated with these tests (p = .69). Two-tailed t-tests were performed on the two pairs of groups intended to be collapsed (i.e., the RM/STAT groups and the TUTOR groups), on CNTL, CONF, and LRT measures and no significant differences were found within the pairs on any of the measures (p >.05), thus lending statistical support to combining the groups accordingly to form the four main conditions.

A one way between subject multivariate analysis of variance was performed on the three dependent variables measuring reasoning: TEXT, SMED, and LRT. The independent variable was training condition. With the use of Pillai's criterion, the combined dependent varables were significantly affected by training condition, $F(9,270) = 2.75$, $p = .004$. In order to investigate more specifically the effect of training on each dependent variable, individual univariate F tests were performed [1]

Gray's Critical Thinking Test (GCAT):

As revealed in Figure 1, on the TEXT measure, subjects in the TUTOR group on average required the least number of cues (M=2.5) in order to recognize the need for comparative information in the text they had just read. Those in the RM/STAT group (M=3.57) did better than the BASELINE group (M=4.13), who, in turn, did better than those in the HUM group (M=4.16). A one way ANOVA revealed a significant training condition effect, $F(3,91)=3.83$, p=.01.

[1] All p values reported are based on two-tailed tests, except for the a priori tests, which are based on one tailed tests.

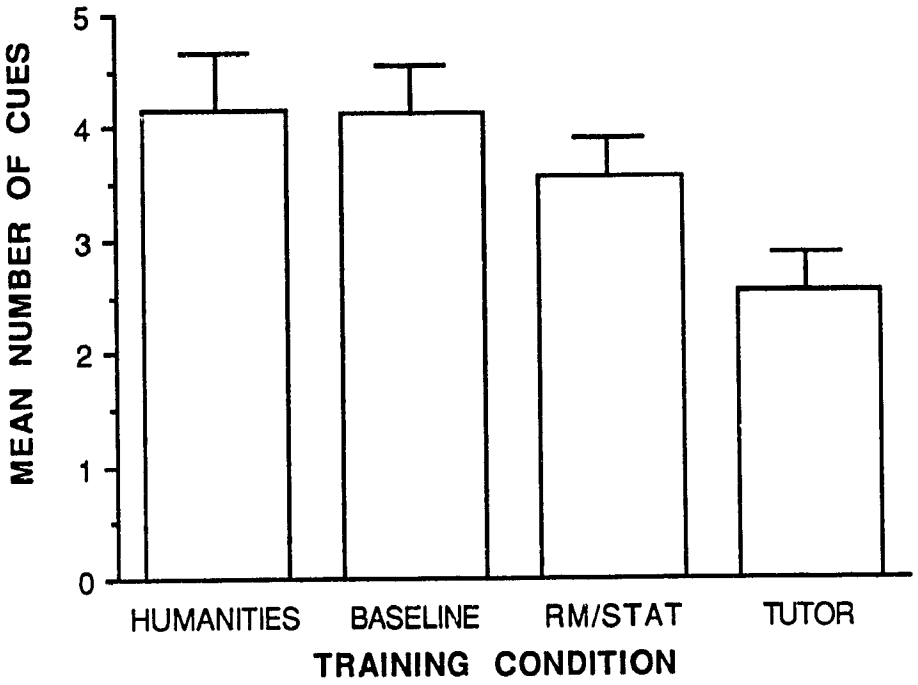


Figure 1. **The mean number of cues needed on the TEXT measure of Gray's Critical Ability Test (GCAT) as a function of training.**

As the main purpose of this study is to assess whether or not the RM/STAT and TUTOR groups' performance on these critical thinking tests differ from that of the BASELINE group, two *a priori* comparisons were performed on the TEXT measure, adjusting significance level to .025 for each individual test. Results indicate that the TUTOR group's performance was significantly better than that of the BASELINE group, $t(46) = 8.81$, $p=.005$. No significant differences were detected between the RM/STAT group and the BASELINE group means on this measure ($p>.05$). As illustrated in Figure 2 , performance on the Smedslund question also improved significantly with training, x^2 (3, N = 95) = 8.85, $p = .03$. Sixty-one percent of the TUTOR group correctly identified the problem with the evidence presented; 46% of the RM/STAT group and 30% of the BASELINE group answered the SMED question correctly; only 21% of the HUM group answered the question suitably.

The Lehman Reasoning Test (LRT):

On average, the TUTOR group correctly answered more questions on the LRT (M = 5.57) as compared to the other three groups (M = 4.30, 4.43, & 4.21 for the BASELINE, RM/STAT & HUM groups, respectively). The ANOVA indicated a significant Training Condition effect, $F(3,91) = 3.26$, $p = .025$. Again, planned comparisons between the TUTOR group and the BASELINE group means were significant, $t(46) = -2.35$, $p =.023$, but not for that between the RM/STAT and BASELINE groups' means, $p >.05$.

The results from the overall LRT significantly correlated with the TEXT measure (r=-.37), and the SMED measure (r=.42), $p < .025$. Thus, better performance on the LRT was correlated with better performance on the TEXT and SMED measures.

In order to better determine which type of questions the subjects were more capable of answering, the questions on the LRT were divided into two categories: (1) knowledge of statistical principles (Stat) and (2) knowledge of methodological principles (Meth). Thus, four questions primarily dealing with the law of large numbers, baserates, or regression towards the mean were categorized in the Stat division, whereas five questions dealing with the influence of confounded variables or the need for control groups or

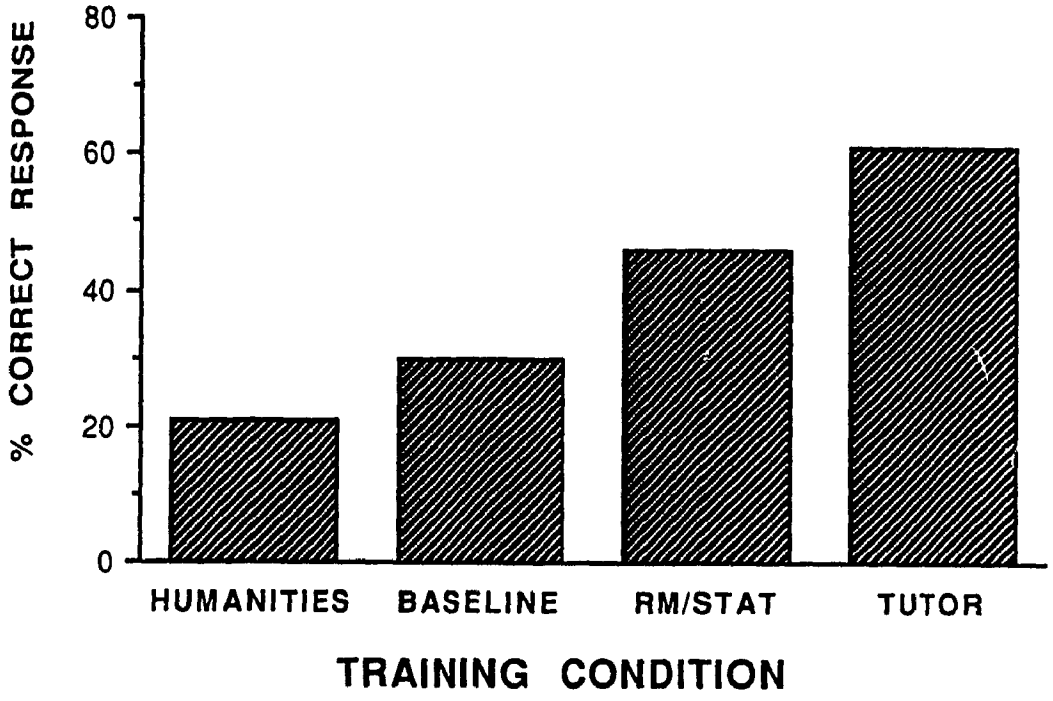


Figure 2. The percentage of respondents correctly answering the Smedslund question as a function of training.

randomization procedures were categorized in the Meth division. The two questions based on the conditional were dropped from analyses for two reasons: (a) they were difficult to classify into either a Stat or Meth group and (b) a floor effect made their contribution to the results negligible. Scores were obtained for each category by simply adding up the number of correct answers each subject made for each category. Figure 3 shows the mean scores across groups on both Meth and Stat categories. The TUTOR group averaged the highest number of questions correctly answered out of the five Meth questions (M=3.04). The HUM, BASELINE, and RM/STAT groups averaged 1.84, 2.20, and 2.21 questions correctly answered, respectively. A two-factor repeated measures ANOVA revealed a significant Main Effect on Training Condition, $F(3,91) = 2.91$, $p = .04$, and a borderline interaction between Training Condition and the Repeated Measure, $F(3,91) = 2.68$, $p = .051$. On the Meth category, planned comparisons between the TUTOR and BASELINE groups were significant, $t(46) = -2.35$, $p = .023$, but not between the RM/STAT and BASELINE groups, $p > .05$. Though the TUTOR group also did the best on the Stat category , there was almost no difference across all group means on the Stat category; means ranged from 2.00 to 2.25.

No significant differences were found across groups in general intellectual functioning, as measured by the Wonderlic Personnel Test, $p = .68$. Scores on the Wonderlic were significantly correlated with the LRT, $r(93) = .35$, $p < .025$, and modestly correlated with the TEXT measure on the GCAT, $r(93) = -.22$, $p < .025$.

Gray's Belief Survey (GBS):

Figure 4 illustrates the overall percentage of participants endorsing belief in each of the ten items in the GBS. Belief was defined as any score falling between +1 and +4. As expected, subjects believed less in the paranormal phenomena than in the more scientifically substantiated phenomena. Participants believed more in ESP (71%) than in Psychic Healing (23%). It is interesting, perhaps, to note that out of all the nonparanormal items, subjects put less faith in Evolution (74%), and believed most strongly that apes can be taught to use sign language (91%).

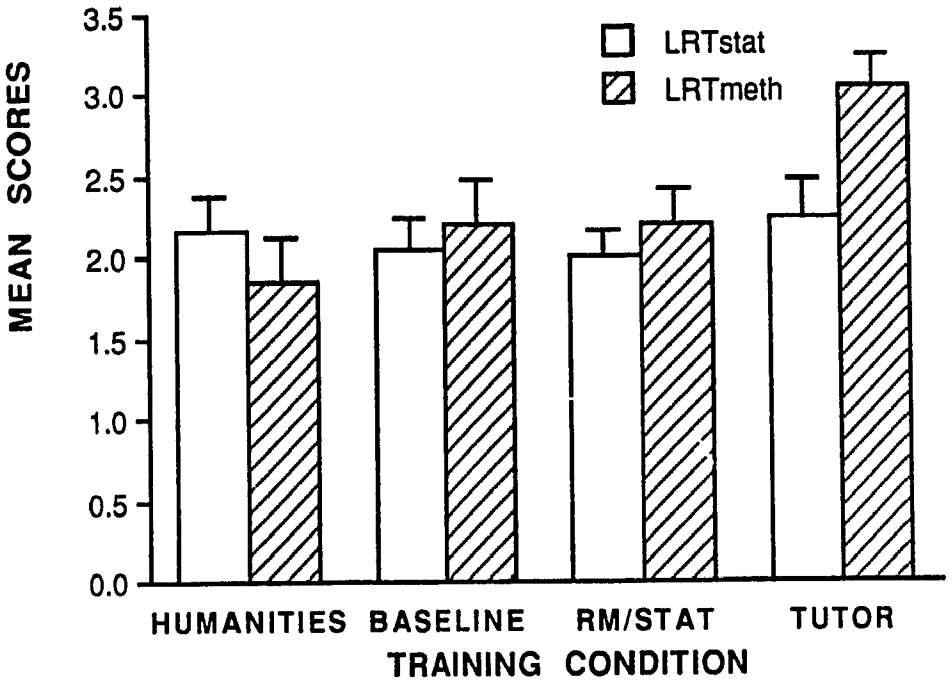


Figure 3. The mean scores on the methodological (Meth) and statistical (Stat) categories of the Lehman's Reasoning Test (LRT), as a function of training.

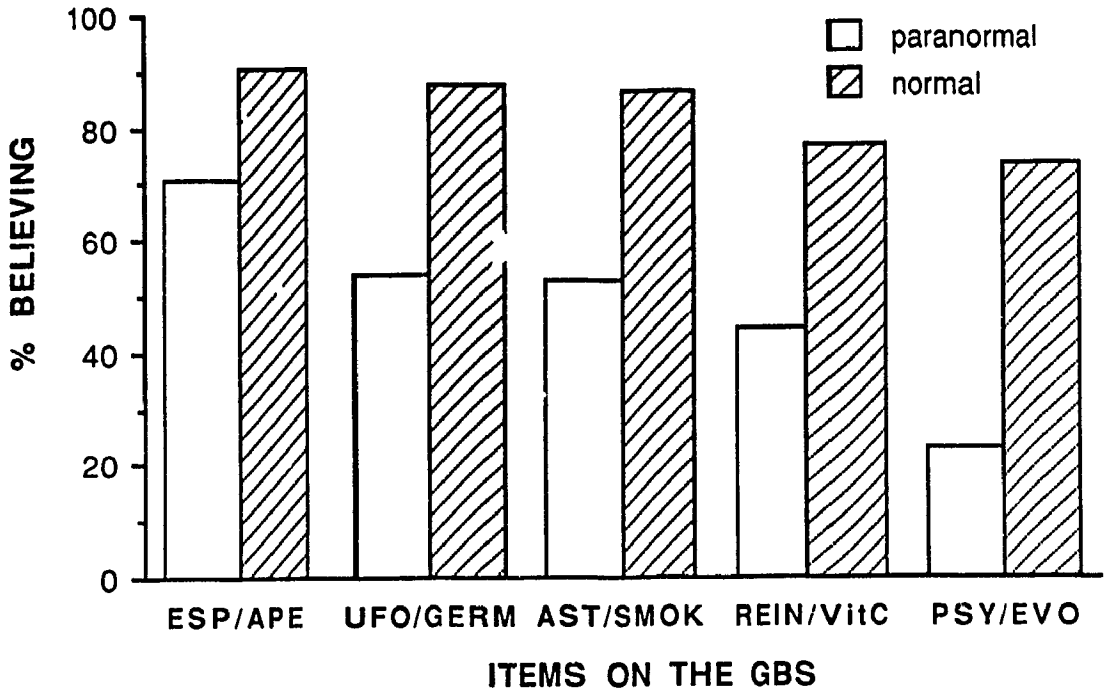


Figure 4. The overall percentage of belief in both normal and paranormal items on the GBS. Note: Items from left to right are: ESP, ape language, UFO, germ theory of disease, astrology, smoking and cancer, reincarnation, vitamin C and colds, psychic healing, and the theory of evolution.

It will be recalled that the TUTOR condition included two groups, namely, one in which students were given training in applying critical thinking skills to normal, everyday-type problems (normTUTOR), and the other in which half of the examples dealt with the paranormal (paraTUTOR). In order to look at the effects of differential training on belief in these unsubstantiated phenomena, the two TUTOR training groups were analyzed separately in order to assess if domain-specific (i.e., paranormal) examples are required in getting students to apply critical thinking skills towards the evidence in the paranormal. Thus, five groups were retained for this part of the analyses: (1) the BASELINE group, (2) the RM/STAT group, (3) the HUM group, (4) the normTUTOR group, and (5) the paraTUTOR group.

An index of overall level of belief in the five paranormal items was calculated by counting the total number of items to which each respondent checked a positive belief score (i.e., +1 to +4). This belief score could, therefore, vary from 0 to 5 depending on how many of the five phenomena the respondents said they believed in. As illustrated in Figure 5, the group with the lowest number of paranormal phenomena endorsed was the paraTUTOR group ($\underline{M}$ = 1.7 items); the HUM subjects believed in the most number of paranormal phenomena ($\underline{M}$ = 3.7 items). The BASELINE, RM/STAT, and normTUTOR groups had mean scores of 2.2, 2.4, and 2.1, respectively. The ANOVA revealed a significant effect on Training Condition, $\underline{F}(4,90) = 5.40$, $\underline{p} = .0006$. Using Scheffé's post-hoc F-test, it was shown that the HUM group differed significantly from all the five psychology/APSS groups , $\underline{p} < .05$. Thus, the HUM group believed in more paranormal phenomena than did the psychology/APSS students. It could be argued that a score of +1 does not fairly reflect a true positive belief, as there was no "0" option available for the unsure believer. Even when defining beliefs as any score between +2 to +4, however, similar results are obtained, in that the overall ANOVA is significant, $\underline{F}(4,90) = 8.13$, $\underline{p} = .0001$, and the HUM group differs significantly from all the other groups ($\underline{p} < .05$) (see Figure 5).

Another way to evaluate overall levels of belief was in terms of respondents' strength of belief. Recall that students indicated strength of belief by checking between −4 through to +4 in the five paranormal phenomena. These scores were averaged to provide a single index of overall belief in the

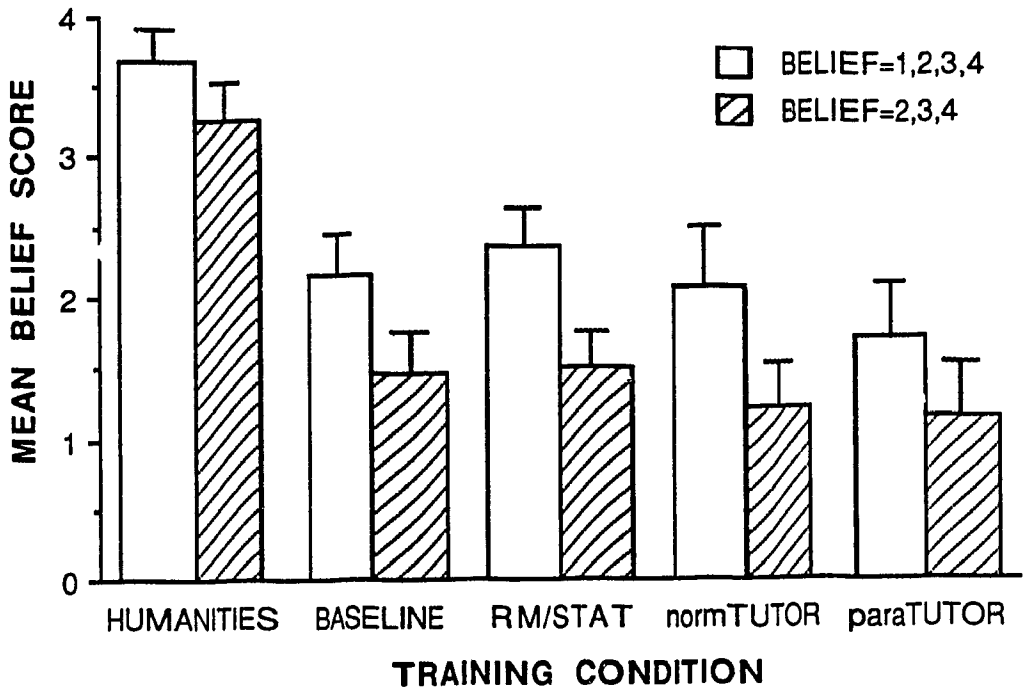


Figure 5. **Mean number of paranormal phenomena in which participants believed, as a function of training. Belief is defined in two ways; as any score falling between +1 to +4, inclusive, or as any score falling between +2 to +4, inclusive.**

paranormal for each respondent. Group means were then calculated from these subject means. It should be noted that a mean strength of belief near zero indicates that belief was roughly evenly divided between belief and disbelief across phenomena. For example, if a respondent gave a +4 score for two items, a -4 score for two items, and a -1 score for one item, this subject would be assigned a strength of belief score of -1. Again, the paraTUTOR group had the lowest level of belief (M= −1.06) and the HUM group had the highest (M= +1.67). The normTUTOR, BASELINE, and RM/STAT groups had means of −.60, −.53, and −.46, respectively. A significant difference was found amongst the groups using a one factor ANOVA, $F(4,90) = 7.82$, $p = .0001$. Scheffé's post-hoc F-test reached statistical significance ($p < .05$) for all the psychology/APSS groups when compared to the HUM group. Thus, the students in the HUM group believed much more strongly in the paranormal than did the psychology/APSS students.

Figure 6 illustrates the breakdown of each paranormal phenomenon across groups. It appears that the reason the HUM group has significantly higher strength of belief scores is due to the high scores in UFO, astrology, and reincarnation.

A final approach in looking at the participants' level of belief in these phenomena was to identify the number of "true" believers and "true" skeptics, defined in this study as those subjects having mean strength of belief scores at the upper and lower 10th percentiles, respectively. As can be seen from the top part of Figure 7, 32% of the HUM subjects were "true" believers, 11% of the RM/STAT subjects were "true" believers, and 5% of the BASELINE group were "true" believers. No "true" believers belonged to the normTUTOR or paraTUTOR groups. The lower part of Figure 7 illustrates the proportion of "true" skeptics within each group. The largest proportion of skeptics belonged, not to the paraTUTOR group (7%) as would be expected, but rather to the normTUTOR group (29%), which was the group predicted to possess the second highest proportion of skeptics. The BASELINE and RM/STAT groups held similar proportions of skeptics (10% and 11%, respectively). No skeptics belonged to the HUM group.

As gender differences have previously been found to play a role in belief in

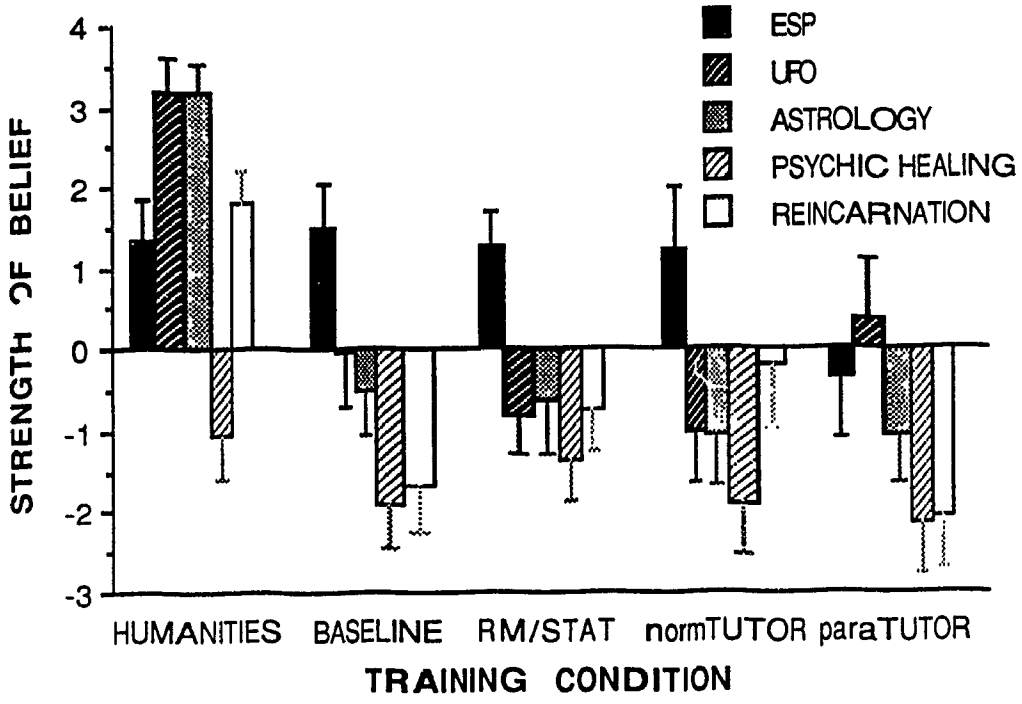


Figure 6. **Overall strength of belief for each paranormal item as a function of training.**

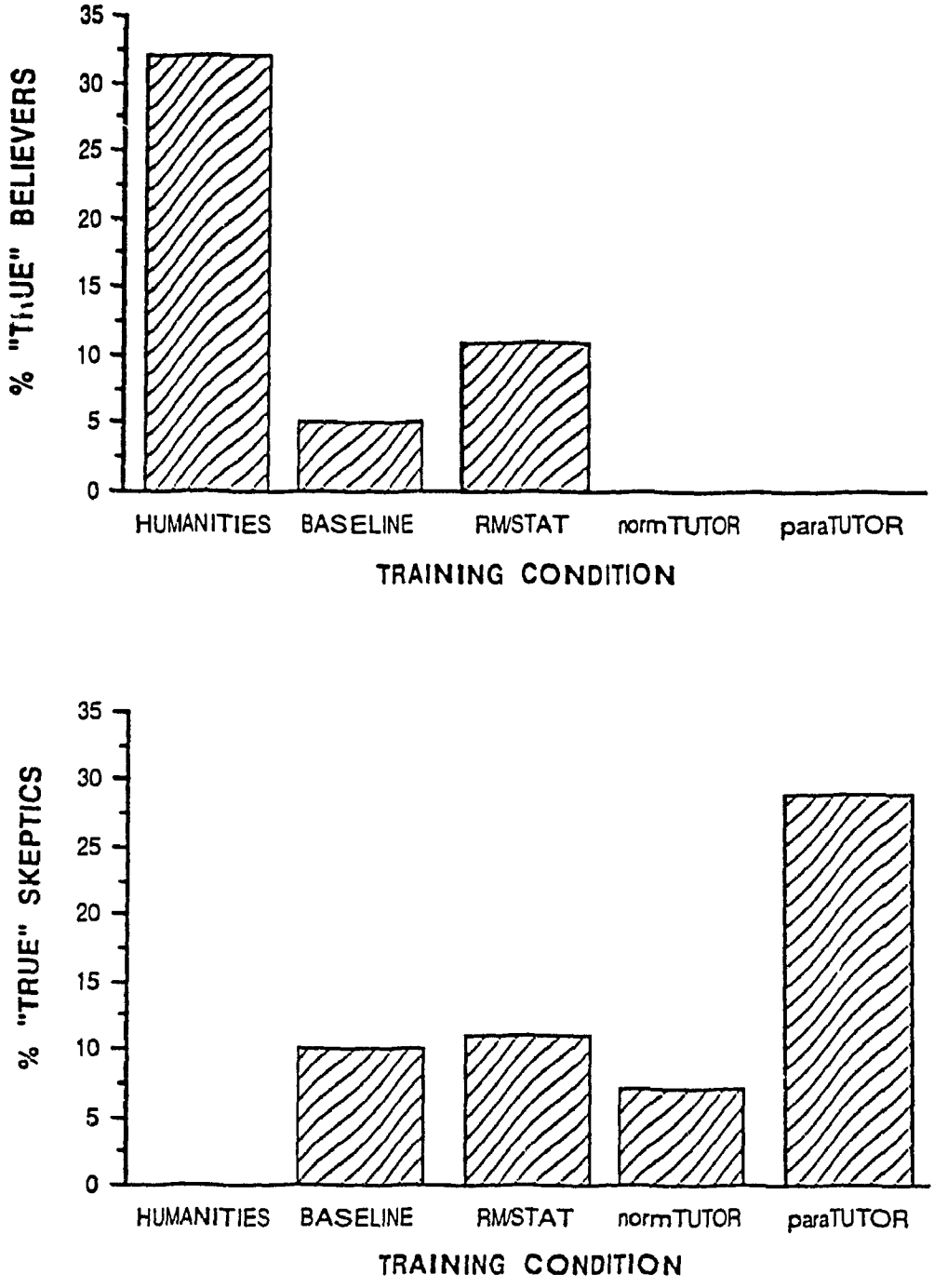


Figure 7. The proportion of subjects within each group either believing (top graph) very strongly (i.e., upper 10th percentile) or disbelieving (bottom graph) very strongly (i.e., lower 10th percentile) in the five paranormal phenomena, averaging strength of belief scores across all phenomena.

unsubstantiated phenomena (Gray, 1990a), it was necessary to see whether or not groups differed with regard to the ratio of men to women. Chi Square analysis revealed no significant differences across groups, χ^2 (5, N = 95) = 3.14, p = .68.

Relationship of Critical Abilities Scores to Belief in Unsubstantiated Phenomena:

The TEXT measure on the GCAT modestly though significantly correlated with the mean strength of belief in paranormal phenomena r(93) = .24, p = .01. That is, respondents with higher levels of belief in the paranormal phenomena needed more cues to recognize the problems with the claims made in the text. Mean strength of belief also correlated with success on the SMED measure, r(93) = -.25, p = .01. This negative correlation is indicative of the fact that success on the SMED measure increases as levels of belief in the paranormal decreases. Finally, the LRT significantly correlated with mean strength of belief, r(93) = -.20, p=.01.

A standard multiple regression was then performed on the overall strength of belief in the paranormal as predicted by the three measures on reasoning (i.e., TEXT, SMED and LRT). The R^2 was found to be significant, F(3,91) = 3.45, p=.02. Altogether, 10% of the variability in strength of belief in the paranormal could be explained by knowing the scores on these three independent variables. In a separate standard multiple regression, the contribution of IQ and SEX in predicting belief in the paranormal was non significant (p>.10), and accounted for less than 2% of the variance.

Discussion

The data presented indicate that the principles learned in psychology research methods (RM) and statistics (STAT) courses only slightly, and not significantly improve students' general statistical and methodological reasoning skills about real-life problems. It is necessary to supplement these courses with tutorial sessions which emphasize the general applicability of these principles

in order to effect significant changes. Humanities subjects performed similarly to the Baseline group.

One interpretation of these findings is that the RM and STAT courses do not emphasize the applicability of scientific principles to other domains; as a result, students are ill-prepared to see how these principles can be generalized to a wide range of issues. Given that one goal of academia is to better equip students with general thinking skills, these results are disappointing. For example, only 3.6% of the RM/STAT subjects spontaneously recognized the need for comparative information on the GCAT. One might have expected that more students would have identified the missing control-group after having spent an entire semester learning the basic principles of the scientific method. As this was not the case, it appears that new approaches to science training that generalize to everyday issues are needed.

While Fong et al. (1986) found that training received in statistics courses was effective in enhancing reasoning performance in everyday problems, the data from the present study only weakly support these findings. That is, the RM/STAT group performed consistently, though not significantly, better than the BASELINE group on these measures, indicating that these courses do have some positive, though only minimal effect on reasoning. Thus far, the results of this study only partially support the formal discipline hypothesis, in that the rules taught in these courses only slightly generalize outside the bounds of psychology to everyday-type problems.

Though an entire semester of RM and STAT courses did very little to enhance methodological-statistical reasoning about everyday problems, the relatively brief tutorial sessions did, in fact, have a significant impact on students' methodological, and to some extent statistical reasoning ability when compared to pre-course levels. As the focus of the tutorial sessions was to demonstrate how basic statistical and methodological principles can, and should, be applied to a variety of everyday situations, it appears that transfer of these principles is possible, as long as the training includes demonstrations of the feasibility of generalization. These findings are consistent with those of Leshowitz (1989) who found that his RM course, which devotes a large segment to teaching inductive reasoning with intensive application to everyday life

events, results in dramatic improvements in general reasoning. Based on these encouraging results, perhaps an emphasis should be routinely placed on how basic reasoning principles taught in the courses can be applied to a wide range of problems. Considering the usefulness of critical thinking in many areas of our lives, as well as the fact that a large proportion of students in university do not necessarily pursue careers in their currently chosen field of study, then statistics and research methods courses could have an important place in the curriculum aside from any usefulness they could have for purely academic issues.

It appears that the notion that reasoning skills are highly domain-specific and subsequently nongeneralizable (e.g., Thorndike, 1905), or else untrainable (e.g., Piaget & Inhelder, 1958) needs to be modified, as results from this study show that transfer can be facilitated by emphasizing the generalizable nature of these principles. As such, the question is not whether transfer is possible, but rather, how teaching methods can best be developed in order to facilitate transfer. This position most resembles that of Nisbett et al. (1983), in that it was shown how the likelihood of applying statistical reasoning to many domains can be improved just by teaching people how to better encode the events to make them more amenable to statistical analysis. In doing so, people can then begin to appreciate that a wider range of problems can be dealt with in a statistical manner. This can explain why the RM and STAT courses did not enhance reasoning skills on problems not directly related to the course material, as the principles taught in the courses are typically presented in a context-specific manner, and students are not encouraged to see the similarities between the examples provided in the classroom to those they may encounter outside the classroom.

It could be argued that perhaps the RM/STAT group was not given enough training in order to promote noticeable changes in critical reasoning. For instance, perhaps a year long course would have resulted in significant improvements. Research has shown, however, that performance even by students who are at the graduate level of study, namely, those having taken a number of RM and STAT courses, is not as impressive as one might hope (Gray, 1990a; Gray & Mill, 1990). An argument against the notion that the formal coursework was too short, is that the special tutorial sessions in this

study resulted in appreciable improvements in reasoning, even though they consisted of only an extra hour of training, which is consistent with the findings of Fong et al. (1986) who also found that very brief training sessions in the law of large numbers increased both the frequency and quality of statistical reasoning as applied to everyday problems. Though amount of training may help improve critical reasoning, seeking out the most efficient and effective approaches to training appear to be more productive, especially for the time-constrained teacher who has a set amount of course material that must get taught in a limited amount of time.

As training in the humanities does not usually include teaching in scientific methods of evidence collection or evaluation, it is not surprising that these students would not excel on tests designed to measure methodological-statistical reasoning skills. The poor performance of the humanities subjects (HUM) on these tests further lends support to the view that differential academic training can influence critical reasoning . These findings are consistent with those of Lehman, Lempert, and Nisbett (1988) who found that graduate training in the probabalistic sciences (e.g., psychology, biology & medicine) helped improve reasoning skills on everyday-type problems, whereas training in the more deterministic science of chemistry did not significantly improve reasoning on these problems over time.

Of course, it is not possible from this study to rule out the potential confound of self-selection; namely, that students with more critical ability enter the more probabalistic/scientific fields to begin with, which could account for why the humanities students performed so poorly. If this is so, it could be argued that academic training may, in fact, play only a small role in the development of a critical thinking attitude, and that perhaps innate ability, early exposure and encouragement to this style of thinking is more important. Note, however, that the humanities students were from the Liberal Arts program, which has a selective admission criteria. Furthermore, it can be counter-argued that academic training can, in fact, be influential in the development of one's reasoning skills, based on the promising results shown by the TUTOR group. Nonetheless, the design of this study was not intended to determine whether or not humanities' education improves critical thinking over time. In order to have assessed that, a humanities baseline group just entering university would have

been required. This study does show, however, that students at similar points in their education differ with regards to their critical thinking skills, depending on their field of study.

It was expected that the TUTOR group would perform the best on both the Meth and Stat subsections of the LRT, and while this prediction was born out, it was surpising to find that the groups did not differ significantly on the Stat subsection. It is possible that the extra tutorial sessions did little to improve statistical reasoning, though this finding is inconsistent with recent reports (e.g., Fong et al., 1986) which suggest that even brief 25 min. training sessions in statistical concepts can improve one's statistical reasoning ability. These sessions, however, focused on specific statistical training (i.e., the LLN), whereas the tutorials from this study placed more emphasis on the need for methodological control (i.e., control groups), which could perhaps explain the differences in the findings. In addition, there were not enough questions on the Stat subsection of the LRT to properly assess statistical reasoning, and two out of the four statistical questions were answered correctly by most subjects; as such, the variability within the scores was too limited to result in significant differences. It is recommended that before using this modified version of the questionnaire in future research, each question should be re-evaluated for its suitability.

The fact that very few subjects were capable of correctly answering the two questions dealing with the conditional on the LRT was not surprising, and is consistent with the findings of Cheng et al. (1985). People do not appear to possess a formalized version of the conditional in their problem solving repertoire. Though their recommendation to teach the material conditional using pragmatic inferential rules was heeded in designing the tutorial sessions, only about 10 min. was devoted to teaching these rules. As these deductive reasoning skills appear a little more difficult to learn, it is probable that more practice than that offered in the tutorial session is required in order to learn these particular rules.

Arguments for the importance of identifying which rules people can most readily be taught are beginning to emerge. For example, Fong et al. (1986) found that the law of large numbers can easily be taught, whereas Cheng et al.

(1985) found it more difficult to teach the material conditional. The results from this study emphasize the relative ease with which people can learn a control-group-way-of-thinking about a wide variety of problems, if given the proper training with plenty of concrete examples. As some clues are emerging about the nature of which rules are more readily integrated into people's problem solving repertoire, perhaps teachers will capitalize on these findings and incorporate new approaches to their teaching methods (if they are not already doing so). Though many teachers find they are pressed for time just covering the required course material, perhaps if they are shown how minimal amounts of time need be invested in order to improve many aspects of students' general reasoning ability, they may be more inclined to incorporate these principles into their course material.

It should be made clear that these tests do not address all aspects of critical thinking ability or reasoning. The GCAT, for example, was designed with the express intention of measuring what could be considered a real-life, everyday-type event (e.g., evidence one might enounter in the daily newspaper) to which a control-group-way-of-thinking could be applied. It in no way covers the multifaceted nature of reasoning but it does address an important methodological concern, namely, whether or not people recognize that comparative information is often needed to properly assess the quality of the evidence.

Though differences between the paraTUTOR and normTUTOR groups on reasoning measures were nonsignificant, the paraTUTOR group typically did slightly worse than the normTUTOR group on most reasoning measures (recall that this group received tutorials on reasoning using many examples from the paranormal). It was the impression of the tutors while giving examples of the paranormal in the paraTUTOR group that scientific explanations were not always welcomed when accounting for these phenomena. In retrospect, it is possible that use of paranormal examples may have inadvertently resulted in a reaction against the critical thinking skills being promoted, or at least detracted from the general efficacy of the basic principles being discussed. This finding may be highlighting the need to carefully consider what types of examples will be most effective in promoting reasoning skills, and which ones may, in fact, be

detrimental to teaching reasoning.

Results from the belief survey (GBS) indicate that training in reasoning has little impact in reducing belief in various scientifically unsubstantiated phenomena. Though it is true that the group receiving reasoning training using paranormal examples (paraTUTOR group) consistently showed the lowest willingness to endorse belief in these phenomena, both in terms of strength of belief and number of phenomena believed in, this group did not differ significantly from any of the other psychology/APSS groups. Furthermore, an entire semester of RM and STAT courses, even when coupled with extra training in reasoning about everyday issues, does not appear to predispose students to reconsider the validity of the evidence supporting the paranormal. This relatively minor reduction in belief in the paraTUTOR group might be explained in terms of the very limited time devoted to training subjects in how to approach the evidence for these phenomena critically. Results, however, show that even an entire semester devoted to debunking these phenomena was only moderately effective in reducing belief in the paranormal (Gray, 1985). Nonetheless, considering how previous research has demonstrated that these beliefs are generally very resistent to change (e.g., Gray, 1985), it is impressive that a total of only one hour of training can have some impact on students' belief in these phenomena. Furthermore, the fact that no "true" believers (i.e., those with levels of belief falling in the upper 10th percentile) were from either the paraTUTOR of normTUTOR groups, and that the largest proportion of "true" skeptics (i.e., those with levels of belief falling in the lower 10th percentile) came from the paraTUTOR group, perhaps indicates that these tutorial sessions may lead some people to be less inclined to endorse strong belief in the phenomena. Of course, whether or not these reductions will remain over time is questionable. As Gray (1985) noted, following an initial reduction in levels of belief in students have partaken in a course designed to challenge belief in these phenomena, belief began to rise again after a one year period.

Though it is clear that training in critical thinking had limited success in reducing belief in the paranormal, the results from the regression analyses suggest that performance on the three reasoning tasks do, in fact, significantly predict overall strength of belief in the five paranormal items. Accounting for

10% of the variance may be, in fact, small, but considering that many factors probably influence these beliefs (e.g., Alcock,1981; Singer & Benassi, 1981), this value should be deemed a relatively substantial finding. Among the many factors that perhaps should not be considered, at least in terms of the present study, are those of IQ and sex, as these factors had a negligible impact in predicting paranormal beliefs. Thus, belief appears to depend less on a person's IQ or sex, and maybe better determined, at least in part, by how critical stance one adopts when evaluating evidence.

It might be argued that the fact that the subjects in the paraTUTOR group had, on average, the lowest levels of belief in the paranormal was actually due to the demand characteristics of the testing situation, namely, they knew that the researchers opposed uncritical belief in these phenomena, and as a consequence, they answered in such a way as to satisfy the experimenter. Though this hypothesis cannot be ruled out, it should be noted that the likelihood of this effect was minimized by having the subjects fill out the GBS in private. Furthermore, it was stressed that the results would not be identified individually.

The most robust finding from the belief survey is that the humanities (HUM) students believed in significantly more phenomena than all the other groups, and that they believed more strongly in these phenomena. This finding is consistent with previous research which also demonstrated that students in nonscientific academic disciplines tend to have higher beliefs in these phenomena, even at the graduate level (Gray, 1990a; Gray & Mill, 1990). Though it is not possible from this study to assert that this is due to differential academic experience because a self-selection confound cannot be ruled out, it does appear that this academic field either tends to attract "believers", or else does not encourage skepticism in these phenomena. Considering the abundance of evidence supporting evolutionary theory, and the controversial nature of the evidence in sup rt of ape language, one surprising finding from the "normal"(i.e., non-paranormal) items on the GBS was that more participants believed that apes could be taught to use sign language than believed in the theory of evolution. This may have been a result of the definition offered for each item, and a review of these definitions may be warranted. Of course, it may just be that people, in general, are less inclined to adhere to the theory of

evolution, and may honestly consider the evidence supporting ape language as more convincing.

The fact that some people believe strongly in some items and not at all in others may attest to the multidimensionality of paranormal belief. In this study, scores from each subject on all the paranormal items were combined to form two indices: total number of items believed in and overall strength of belief. In combining all the items together to obtain one overall score, however, valuable information may have been lost. Perhaps looking at each paranormal phenomena individually, and investigating which groups of people are believing in specific phenomena may help clarify what factors are involved in these beliefs. For example, are there particular groups of people believing in certain phenomena? As a case in point, Gray (1990) generally found that more males believe in UFOs and that more females believed in ESP. As such, in order to better understand the basis of beliefs in the paranormal, it may be worthwhile to treat each phenomenon separately.

Based on the findings from this and other studies, it would be productive to develop either a course on enhancing general reasoning skills that can be applied to everyday experiences, or develop simple, relatively straightforward approaches teachers can use when teaching their own course material, in order to extend the scope of the rules they teach. As such, it may be helpful to investigate what effects the tutorial sessions would have on their own, independent of the RM and STAT courses. It would also be interesting to see if these enhanced reasoning approaches have any effect on students' ability to learn the course material, as there is the possibility that learning how the rules apply to everyday situations may have the counter-effect of improving understanding of the rules within the course context. Unfortunately, consent to look at individual course grades was not obtained for this study, so these analyses, for the time being, cannot be done. Finally, revising some of the reasoning tests so that all subjects could be pre- and post-tested would lend more power to the design by reducing any differences that may exist between the groups. For example, though psychology/APSS subjects were randomly assigned to each condition and no significant differences were found across groups on a variety of variables, the groups were not perfectly matched on

important variables such as intelligence and gender. By employing a pre-post test, all subjects could serve as their own control, thus reducing the potential for error.

References

Alcock, J. (1981). Parapsychology: Science or magic? New York: Pergamon.

Alcock, J., & Otis, L. (1980). Critical thinking and belief in the paranormal. Psychological Reports, 46, 479-482.

Allan, L. G., & Jenkins, H. M. (1980). The judgement of contingency and the nature of the response alternative. Canadian Journal of Psychology, 34, 1-11.

Alloy , L.B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. Psychological Review, 91, 112-149.

Chapman, L.J., & Chapman, J.P. (1967). Genesis of popular but erroneous psycho-diagnostic observations. Journal of Abnormal Psychology, 72(3), 193-204.

Cheng,P.W., & Holyoak, K.J. (1985). Pragmatic reasoning schemas. Cognitive Psychology, 17, 391-416.

Cheng, P. W. , Holyoak, K.J. , Nisbett, R. E., & Oliver, L.M. (1986).Pragmatic versus syntactic approaches to training deductive reasoning. Cognitive Psychology, 18, 293-328.

Dodrill, C.B. (1983). Long term reliability of the Wonderlic Personnel Test. Journal of Consulting and Clinical Psychology, 51(2), 316-317.

Fong,G.T., Krantz, D.H. , & Nisbett, R.E. (1986). The effects of statistical training on thinking about everyday problems. Cognitive Psychology, 18, 253-292.

Galotti, K.M. (1989). Approaches to studying formal and everyday reasoning. Psychological Bulletin, 105(3), 331-351.

Gray, T. (1985). Changing unsubstantiated belief: Testing the ignorance hypothesis. Canadian Journal of Behavioural Science, 17, 263-270.

Gray, T. (1987). Educational experience and belief in the paranormal. In F. Harrold & R. Eve, (Eds.), Cult Archaeology and Creationism : Understanding Pseudoscientific Beliefs about the Past. University of Iowa Press.

Gray, T. (1990a). Gender differences in belief in scientifically unsubstantiated phenomena. Canadian Journal of Behavioural Science , 22, 181-190.

Gray, T. (1990b). Questionnaire format and item context affect level of belief in both scientifically unsubstantiated and substantiated phenomena. Canadian Journal of Behavioural Science, 22(2), 173-180.

Gray , T. (in press). Short supplementary report: Critical abilities, graduate education (Psychology) and belief in unsubstantiated phenomena.

Gray, T., & Mill, D. (1990). Critical abilities, graduate education (Biology versus English), and belief in unsubstantiated phenomena. Canadian Journal of Behavioural Science , 22(2), 162-172.

Greenwald, A.G., Pratkanis, A.R. , Lieppe,M.R. , & Baumgardner, M.H. (1986). Under what conditions does theory obstruct research progress? Psychological Review, 93(2),216-229.

Harrold, F., & Eve, R. (Eds.). (1987). Cult Archaeology and Creationism: Understanding Pseudoscientific Beliefs about the Past. University of Iowa Press.

Inhelder, B., & Piaget, J. (1958). The Growth of Logical Thinking from Childhood to Adolescence. New York: Basic Books.

Jenkins, H. M., & Ward, W. (1965). Judgement of contingency between responses and outcomes. Psychological Monographs, 79(1).

Jones, W.H., & Russell, D. (1980). The selective processing of belief disconfirming information. European Journal of Social Psychology, 10, 309-312.

Kahneman,D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3, 430-454

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80, 237-251.

Kahneman, D., Slovic, P. & Tversky, A. (1982). Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: formal discipline and thinking about everyday-life events. American Psychologist, 43, 431-442.

Lehman, D. R., & Nisbett, R. E. (in press). A longitudinal study of the effects of undergraduate training on reasoning. Developmental Psychology, 26 (6),

Leshowitz, B. (1989). It is time we did something about scientific illiteracy. American Psychologist, 8, 1159-160.

Lord, C.G., Ross, L. & Lepper, M.R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology, 37(11), 2098-2109.

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching Reasoning. Science, 238, 625-631.

Nisbett, R. E., Krantz, D., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. Psychological Review, 90, 339-363.

Nisbett, R., & Ross, L. (1980). Human Inference: Strategies and Shortcomings of Social Judgment. Englewood Cliffs, NJ: Prentice-Hall.

Otis, L.P., & Alcock, J.E. (1982). Factors affecting extraordinary belief. The Journal of Social Psychology, 118, 77-85.

Pascarella, E.T. (1989). The development of critical thinking: Does college make a difference? Journal of College Student Development, 20(1), 19-26.

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. Journal of Educational Psychology, 77, 562-571.

Russell, D. & Jones, W.H (1980). When superstition fails: Reactions to disconfirmation of paranormal beliefs. Personality and Social Psychology Bulletin, 6 (1), 83-88.

Singer, B., & Benassi, V.A. (1981). Occult beliefs. American Scientist, 69, 49-55.

Smedslund, J. (1963). The concept of correlation in adults. Scandinavian Journal of Psychology, 4, 165-173.

Sternberg, R.J., Smith, E.E. (1988).The Psychology of Human Thought. Cambridge University Press.

Tversky, A., & Kahneman, D. (1973). Availabilty: A heuristic for judging frequency and probability. Cognitive Psychology, 5, 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science,185, 1124-1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. Psychological Review, 90, 293-315.

Wierzbicki, M. (1984). Reasoning errors and belief in the paranormal. The Journal of Social Psychology, 125(4), 489-494.

Wonderlic, E.F. (1983). The Wonderlic Personnel Test - Form A. U.S.A.

APPENDIX I

RECRUITMENT SHEET FOR THE PSYCHOLOGY / APSS STUDENTS

PARTICIPATION IN A RESEARCH PROJECT

We would like your help with some interesting research that we are doing concerning how people deal with the vast amounts of information they receive daily. One of the things we are interested in is concerned with peoples' abilities to estimate the trustworthiness or reliability of the information they are given by newspapers, TV, friends, (or even professors).

Your participation would involve answering some questionnaires and some interesting tests. Your participation is voluntary and will not count towards your grade in the course. However, you will probably find that your experience in the research project will be valuable to you in connection with your research methods and statistics course as well as your other courses and in everyday life.

The basic testing session will take about 1 hour and 15 mins, but some participants will be asked to come for 2 or 3 other very brief (15 min) sessions during the term.

The procedures are painless, and do not involve anything embarrassing or dangerous.In fact you will probably find the tests quite interesting. All results will be kept absolutely confidential.

PRIZES: We are offering an opportunity to win a $100.00 prize!

You will have a chance to win one of **3,** $100.00 prizes. The draw for the prizes will take place at the end of the semester. [Your chances of being a winner (as you will find out in your stats class) are very much better than in the LOTTO!]

You will notice that this information is printed on **different coloured paper for different students.** We would like you to save this page because we will be making appointments for you depending on the colour you have been randomly assigned (see, you are already finding out about research methods.)

REMEMBER YOUR COLOUR! It is important that you do not exchange your colour.

NB. For today we would like all students with a **YELLOW** page to meet with us for 2 minutes after class to choose a convenient time to come for the test session next week. We will make arrangements on **another occasion** for students with **other colours.**

The research is part of an on-going project supervised by Dr. T. Gray, and it will be carried out by Davina Mill (graduate student), Pauline Kafka and Christine Lavoie (Research Assistants), and David Stone (Honours student).

The testing will take place in H663-1 or H661-2. A contact phone number is 848-2211 (Dr. Gray)

APPENDIX II

PERSONAL DATA FORM

Personal Data

1. ID # :

2. Sex : ___male female___

3. Age : ____years old

4. Did you go to CEGEP? ___yes no___

If yes, what was your program of study?

____social science
____health science or pure and applied
____other (please specify) _____

5. What degree are you going for at Concordia? (e.g. major in English;first year)

_____

6. Was your high school and CEGEP in english? ___yes no___

7. Is this your first university degree ? ____yes no____

(If you answered "yes" to #7, go to question #8)

If you answered "no" to the above question:

a) In what was your previous degree? (eg biology)

b) How many years of study did you complete in this area?
_____ years

8. Did you pass CEGEP Math 337 (statistics)?

____yes
____no

APPENDIX III

GRAY BELIEF SURVEY (GBS)

We are asking you to respond to a brief questionnaire concerning belief in various phenomena and ideas.

Your participation is completely voluntary and your responses will be kept confidential. It will only take a few minutes for you to give your response, but we want you to read a brief description of what we mean by each of the 10 items on the questionnaire form.

Please remember that we are asking if you believe in the reality of the item not just the theoretical possibility. That is, do you feel that the reality of the phenomena has been demonstrated.

ESP We are asking whether or not you believe in ESP (extrasensory perception). Telepathy (thought transference or mind reading) or clairvoyance (the ability to be aware of events that are not in sight) are typical examples of ESP. In general, ESP involves the ability to obtain information without using our senses.

EVOLUTION We are asking whether you believe that humans are the products of an evolutionary process whereby all organisms have arisen by descent from a common ancestor.

"GERMS" We are asking if you believe in the "germ theory" of disease." Germs" is used here as short for the notion that diseases are caused by bacteria or viruses.

UFOs We are asking whether or not you believe in UFOs (unidentified flying objects).That is, do you believe that the earth is visited by,or has been visited by spacecraft of extraterrestrial origin (from outer space).

APE LANGUAGE We are asking whether or not you believe that apes (like Koko the gorilla) and chimpanzees (like Washoe) can talk with sign language. The claim has been made that they can be taught to use signs as a real language like human language.

ASTROLOGY We are asking whether or not you believe in astrology. Briefly, astrology involves the belief that the position of the stars and the planets at the time of your birth affects your personality or what happens to you.

PSYCHIC HEALING We are asking whether or not you believe in psychic healing of medical problems, e.g. healing by the "laying on of hands". We are thinking here of actual cures that are brought about by the special powers of the psychic healer.

SMOKING AND CANCER We are asking whether or not you believe that smoking increases ones chances of getting cancer.

REINCARNATION We are asking whether or not you believe in reincarnation. By reincarnation we mean, for example, that humans can come back, after death, as other humans.

VITAMIN C AND COLDS We are asking whether or not you believe in the claim that high doses of vitamin C can prevent and/or cure colds.

Please fill out the attached questionnaire form. **Remember that we are interested in whether you believe in the reality of each item, not just the theoretical possibility.**

On the form below simply mark the appropriate box indicating YES or NO depending on whether you believe or not. Please note that you indicate the strength of your "YES" or "NO" by checking the appropriate box --- 1 = Weak 4 = Strong.

No					Yes			
4	3	2	1		1	2	3	4
				ESP				
				EVOLUTION				
				"GERMS"				
				UFOs				
				APE LANGUAGE				
				ASTROLOGY				
				PSYCHIC HEALING				
				SMOKING and CANCER				
				REINCARNATION				
				VITAMIN C and COLDS				

APPENDIX IV

LEHMAN REASONING TEST

INSTRUCTIONS

Please read each question completely and indicate what you feel is the best answer by circling the correct letter that corresponds to your answer. Take your time and think carefully.

Your answers will be kept confidential.

Please do not communicate with other people who may be in the room with you. Please do not ask the experimenter any questions. No more information than is already in the question can be given. If you have any comments you could make a note on the back of your answer sheet.

Please make sure you answer every question to the best of your ability.

A researcher at the University of Pennsylvania recently concluded that African violet plants don't like to be yelled at. The researcher, a physics professor, and some of his students discovered this in an experiment with these plants. In the experiment a student transplanted two identical African violet plants from the same greenhouse and grew them under identical conditions, except that the first plant that was transplanted was exposed to about 100 decibels of noise (approximately the same as a person would hear while standing on a busy subway platform) while the African violet transplanted second was grown in quiet conditions. After 1.5 weeks of continuous exposure, the sound-treated plant but not the other wilted.

Assuming that the phrase "same greenhouse conditions" means that every effort was made to treat the two plants exactly the same, which one of the following changes would have done the most to improve the quality of the experiment and the credibility of any results suggesting that noise harms plants?

a) A coin should have been flipped to determine which plant received the noise rather than determining this by which was first transplanted.

b) Ten African violet plants should have been used in each of the two conditions.

c) When the first plant died it should have been studied for evidence of insect damage or fungus.

d) Because houseplants (like African violets) are ordinarily exposed to low levels of noise, outdoor plants that are ordinarily exposed to more noise should have been used.

The city of Middleopolis has had an unpopular police chief for a year and a half. He is a political appointee who is a friend of the mayor, and he had little previous experience in police administration when appointed. The mayor has recently defended the chief in public, announcing that in the time since the chief took office, crime rates decreased by 12%. Which of the following pieces of evidence would most <u>weaken</u> the mayor's claim that his chief is competent?

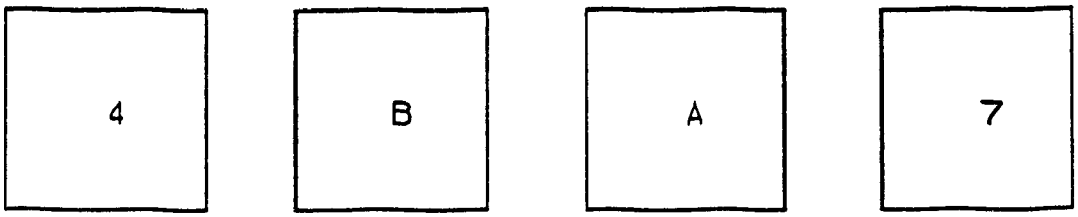
a) The crime rate of the two cities closest to Middleopolis in location and size have decreased by 18% in the same period.

b) An independent survey of the citizens of Middleopolis shows that 40% more crime is reported by respondents in the survey than is reported in police records.

c) Common sense indicates that there is little a police chief can do to lower crime rates. These are for the most part due to social and economic conditions beyond the control of officials.

d) The police chief has been discovered to have business contacts with people who are known to be involved in organized crime.

Below are four cards. They are randomly chosen from a deck of cards in which every card has a letter on one side and a number on the other side. Your task is to say which of the cards you need to turn over in order to find out whether the following rule is true or false. The rule is: " If a card has an 'A' on one side, then it has a 4 on the other side." Turn over only those cards that you need to check the rule.



Box 1 Box 2 Box 3 Box 4

a) Box 3 only

b) Boxes 1, 2, 3, and 4

c) Boxes 3 and 4

d) Boxes 1, 3, and 4

e) Boxes 1 and 3

After the first two weeks of the major baseball season, newspapers begin to print the top ten batting averages. Typically, the leading batter after two weeks has an average of about .450. Yet no batter in major league history has ever averaged as high as, or even close to, .450 at the end of the season. What is the most likely reason for this?

a) A player's high average at the beginning of the season may be just a lucky fluke.

b) A batter who has such a hot streak at the beginning of the season is under a lot of stress to maintain his performance record. Such stress adversely affects his playing.

c) Pitchers tend to get better over the course of the season as they get more into shape. As pitchers improve, they are more likely to strike out batters, so batters' scores go down.

d) When a batter is known to be hitting for a high average, he stops getting good pitches to hit. Instead, pitchers "play the corners" of the plate since they don't mind walking him or else they bear down more when they pitch to him.

The promoters of a local Weight Watchers organization have claimed that, on the average, their members lose fifteen pounds during their first three months of attending meetings. To test this claim, a public health nurse kept records of weight lost by every new member who joined the Weight Watchers branch during 1986-1987. Out of 138 people who started to attend the meetings, 81 kept attending for at least three months. And indeed, the average amount of weight lost by these people was 14.9 pounds.

Based on these facts, what is the best assessment we can give of the likely success of the program?

a) The techniques used in the program are likely to help most people who enroll.

b) The techniques used in the program are likely to help half or fewer of the people who enroll.

c) Those who enroll in the program and stick to it will on the average be better off at the end, although it is impossible to say whether the techniques used in the program are effective.

d) Those who enroll in the program and stick to it will on the average be better off at the end, because the techniques used in the program are effective.

"New! Grapefruit tablets for the most effective diet ever. Watch pounds roll off! (Warning: Dieters must not restrict their food to the tablets alone, but must eat no more than three balanced meals a day.)"

Which of the following is the strongest criticism of the claim made in the advertisement that the tablets will cause weight loss?

a) It is not sensible to expect weight loss to result from the consumption of nutrients.

b) If the tablets are used as directed and weight loss occurs, the restriction of food intake to three balanced meals could be the cause.

c) Weight loss is normally achieved primarily by a combination of exercise, proper nutrition and dieting.

d) Although the advertisement says, "Watch pounds roll off!", it does not specify when weight loss will occur.

e) The amount of weight lost on such a diet depends entirely on the amount the dieter is overweight at the beginning of the diet.

Susan began smoking when she was 18 years old. Even though she read that most experts in the field agree that smokers have a 75% greater chance than non-smokers of developing some form of cancer, she's not worried. Instead, she'll direct your attention to another article that she read about a man who smoked a pack of cigarettes every day for 60 years, and lived to the age of 105! Based on this, Susan has not taken any action to quit smoking.

Which of the following comments most accurately explains why the old man lived so long, even with the odds stacked against him?

a) Everyone is unique, which makes calculating probabilities on how long people will live a ridiculous endeavor in the first place.

b) It is probable that the old man's longevity was an exceptional case.

c) The old man was probably rich enough to afford the best health care.

d) The old man was not worried about his smoking habit, which consequently reduced his stress level, and thus increased his lifespan.

e) The old man heeded the surgeon general's warning, and did not inhale when he smoked.

Tests done on the residents of an urban neighbourhood built on a site formerly used as a dump for chemical wastes showed that nearly 15% of the adults had abnormal chromosome patterns. Abnormal chromosome patterns can be caused by radiation, chemical fumes, and other impurities in the air.

Which of the following would be the most useful information in determining whether the residual effects of the chemical wastes are responsible for the abnormal chromosome development?

a) Whether the abnormal patterns are permanent or reversible.

b) Whether only certain types of people in the area had developed the abnormal chromosome patterns.

c) If over 15% of the people living in another chemical dump area had developed abnormal chromosome patterns.

d) What the long-range effects of these abnormal chromosome patterns will be.

e) Whether abnormal chromosome patterns are found in other urban adults.

Smith says: "During the late 18th century in the USA, the Industrial Revolution created large numbers of jobs in manufacturing centers, causing a sudden shift in population density from the rural South to the industrializing North".

Jones says: " Your explanation is an oversimplification. It was also the destruction of the existing system of agriculture that drove people into the cities".

Which of the following additional pieces of data would support the Jones conclusion?

I. Immediately prior to the sudden population shift, newly introduced principles of scientific agriculture markedly improved farm efficiency.

II. The population shift began some twenty years before the construction of the first factories.

III. At the same time the Industrial Revolution was beginning, the courts declared that common areas owned by municipalities and used by farmers for grazing stock could be sold to private individuals.

a) I and II

b) I and III

c) II and III

d) III only

e) All of the above

A talent scout for a professional symphony orchestra attends a musical competition with the intention of observing carefully the talent and skill of a particular violinist. In each of the first six performances, the violinist repeatedly plays difficult passages with a fluency worthy of the best professional performers. However, in the final performance of the competition, as one of the two semi-finalists, the player stumbles over a key solo passage, stops playing, and tries again. The other semi-finalist performs her concerto flawlessly, and goes on to win the competition.

The scout reports that the player in question "has excellent skills, and should be recruited. He has a tendency to misplay under extreme pressure, but this will probably disappear with more experience and training".

The scout's report is:

a) probably accurate both in assessing the player's general level of ability and his tendency to misplay under pressure.

b) probably accurate in assessing the player's general level of ability, but perhaps inaccurate in assessing his tendency to misplay under pressure.

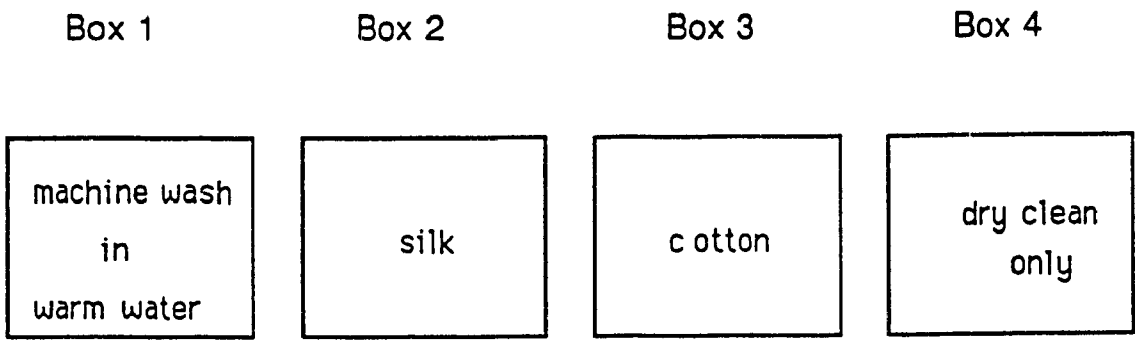
c) perhaps inaccurate in assessing the player's general level of ability, but probably accurate in assessing his tendency to misplay under pressure.

d) perhaps inaccurate in assessing both the player's general level of ability and in assessing his tendency to misplay under pressure.

e) probably inaccurate in assessing both the player's general level of ability and in assessing his tendency to misplay under pressure.

As part of your job as a quality control inspector at a shirt factory, you have the task of checking fabric and washing instruction labels to make sure that they are correctly paired. Fabric and washing instruction labels are sewn back to back. Your task is to make sure that all "silk" labels have the "dry-cleaning only" label on the other side.

Which of the following labels would you need to turn over? Turn over only those you would have to check to be sure.



a) Box 1 only

b) Boxes 1 and 2

c) Boxes 2 and 4

d) Box 2 only

e) All of the above