Exploring the Mechanisms Underlying Gender Differences in Statistical Reasoning: A Multipronged Approach

by

Nadia Martin

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Psychology

Waterloo, Ontario, Canada, 2013

Author's Declaration page

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The past two decades have seen a substantial increase in the availability of numerical data that individuals are faced with on a daily basis. In addition, research uncovering the multiple facets of statistical reasoning has become increasingly prominent. Both gender differences and the effect of experience or training have emerged as two key factors that influence performance in statistics. Surprisingly, though, the combined effects of these two variables have not been studied. This gap in understanding the joint effect of gender and experience on statistical reasoning is addressed in the present dissertation with six studies. In Study 1 (N = 201), participants with various levels of experience in statistics were asked to complete the Statistical Reasoning Assessment (SRA; Garfield, 2003). Although the performance of both genders improved with experience, the gender gap persisted across all experience levels. Multiple measures of individual differences were used in a confirmatory structural equation model. This model supported the idea that differences in statistical reasoning are not uniquely a matter of cognitive ability. In fact, gender was found to influence statistical reasoning directly, as well as indirectly through its influence on thinking dispositions. In Studies 2 (N = 67), 3 (N = 157), and 4 (N = 206), the role of stereotype threat was examined as a potential cause of the persisting gender gap in statistics, and value affirmation was tested as an intervention to overcome stereotype threat. Despite the fact that many women believed negative stereotypes about the ability of women in statistics, value affirmation had no significant impact on performance. To help explain this lack of effect, and in keeping with the results of the structural equation model suggesting a multi-pronged approach, efforts were turned towards a different (and potentially richer) cognitive factor. Specifically, mental representations were explored to help shed light on the root causes of those conceptual understanding differences in statistics. In Studies 5 and 6, gender differences in mental representations of statistical features were examined using a categorization paradigm. In Study 5 (N = 219), extending some of the key findings in Studies 1, 3 and 4, it was established that two courses in statistics are necessary to create a significant difference in the quality of mental representations of statistical concepts. More importantly, Study 6 (N = 208) demonstrated how constraining the task format particularly benefits women in that the quality of their reasoning significantly improved, where that of men was equal across tasks. Theoretical and practical implications of these findings are discussed.

Acknowledgements

As my journey as a graduate student at the University of Waterloo is coming to an end, I would like to take this opportunity to thank the many great souls that made this accomplishment possible. Whether you have contributed by challenging me intellectually, by extending your friendship and emotional support to me, or by offering yourself as a role model, please know that I never could have made it so far and so successfully without you on my path.

More particularly, I would like to thank Dr. Jonathan Fugelsang for his relentless energy, outstanding patience, and constant support. You are truly one of a kind!

I would also like to thank Dr. Derek Koehler for trusting me and involving me in his teaching. My experience teaching statistics has been invaluable as a stepping-stone to finding my way to a dissertation and has been key to keeping me motivated throughout the process.

It is also important that I acknowledge my team of readers and reviewers: Dr. Bob delMas, Dr. Erik Woody, Dr. Mike Dixon, and Dr. Steve Mok, all of whom have dedicated precious time to helping me improve my dissertation, thus allowing me to reach ever slightly further.

Krystelle, your call for a writing group was pure serendipity. You made me accountable at the time it mattered most and made the writing of this thesis possible. I wonder how I would have managed without you. As such, I am forever grateful!

Peter, thank you for accompanying me throughout this journey, sacrificing many exciting adventures to come sit with me at a café while I did my work, as well as for letting me use your dishwasher to let me regain control over my life more than once!

Finally, I want to thank my parents. Without their willingness to invest in my education early on, letting me move to cities increasingly further from home, and without the work ethics they modeled throughout their life, I never would have made it all the way here.

Table of Contents

List of Figures

List of Tables

GENERAL INTRODUCTION

Statistical thinking will one day be as necessary a qualification for efficient citizenship as the ability to read and write. --S.S. Wilks

With the increasing amounts of numerical information that permeate modern life, work and civic life demand citizens to have at least some degree of statistical literacy (Wallman, 1993; Ben-Zvi & Garfield, 2008). The new door to knowledge is data (Lohr, 2009) and statistical competence holds the key to that door. Indeed, we know that statistical competence – numeracy paired with critical thinking – allows for proper evaluation of data to guide decision- and policy-making. In contrast, a lack of such competence is not only disadvantageous, but can have undesirable effects and create ethical dilemmas, as when consent is given to take a new prescription drug despite a lack of proper understanding of the risks involved (e.g., Couper & Singer, 2009; Reyna, Nelson, Han, & Dieckmann, 2009; McHugh & Behar, 2009). Today, a low level of numeracy is detrimental to informed decision-making (e.g., choosing between two medical treatments), and to employability, with outcomes potentially worse for women than men (Parsons & Bynner, 1997, 2005).

The new faces of work and access to information have already influenced the structure of education in the field of statistics. Notably, a new statistics education curriculum has been adopted in the United States in the recent past. The Guidelines for Assessment and Instruction in Statistics Education (GAISE) (ASA, 2005) are the culmination of a long process initiated by George Cobb in 1992, where he recommended emphasizing statistical thinking, in addition to focusing on data and concepts rather than calculations. Thirteen years later, GAISE (ASA, 2005) reprised those recommendations in setting its guidelines for a first course in statistics at the college level. In particular, the report stresses the importance of developing statistical literacy and statistical thinking, and advocates conceptual understanding over mere procedural knowledge. Some of the other accepted learning goals in this new era of statistical education (Garfield & Gal, 1999) include understanding the purpose and logic of statistical investigations, learning statistical skills such as organizing data and constructing tables, developing useful statistical dispositions such as demonstrating critical reasoning when assessing evidence, as well as developing statistical reasoning – the ability to make sense of statistical information. Stressing conceptual understanding makes sense at a time when computer tools and software packages can easily handle all calculations. The question is: how well are we equipped to assess statistical conceptual understanding?

Assessing statistical competence

With the introduction of the GAISE report (ASA, 2005) came the need to measure the impact of GAISE recommendations on students' learning. Assessment of statistical competence, as defined by the new curriculum, became an important goal for educators and researchers in the field of statistics. One instrument allows the accomplishment of this goal - the Statistical Reasoning Assessment (SRA) (Garfield, 1991, 1998, 2003; Garfield & Gal, 1999). The SRA allows educators to measure development and achievement in the classroom, while its ease of scoring provides an accessible tool to instructors of large classes and to researchers (Garfield & Chance, 2000). Designed to assess a wide range of statistical concepts covered in high school and in introduction to statistics classes at the college level, the SRA has the particular advantage of measuring both correct reasoning – such as distinguishing between discrete versus continuous data, understanding the nature of samples and the measures used to describe them, and reasoning about uncertainty and randomness – and misconceptions. Going beyond simple incorrect reasoning, statistical misconceptions reflect beliefs, interpretations or understandings that are mistaken (but often intuitively plausible). Such misconceptions can be resistant to change (Chi & Roscoe, 2002) and impervious to instruction (Konold, 1995). Examples of misconceptions in statistics include thinking that groups cannot be compared if they are not the same size, failing to take outliers into consideration when computing the mean, judging probabilities based on representativeness, and assuming that small samples are as good as large ones for drawing conclusions. Despite their intuitive appeal, those misconceptions are at odds with a technical understanding of statistical principles. For example, in spite of the fact that larger samples improve prediction, many people still trust small samples to be representative of the population (Kahneman, Slovic, & Tversky, 1982) and base their decisions on them.

The inclusion of many different areas of understanding within one single research tool breaks from the tradition of much published research. It is common to read articles focusing on a single aspect of statistical reasoning such as the law of large numbers (e.g., Fong, Krantz, and Nisbett, 1986), the need for comparison groups (e.g., Gray & Mill, 1990), or the importance of base rates in probability judgments (e.g., Bar-Hillel, 1980). Although the inclusion of a range of topics in the SRA makes for a relatively low internal consistency, its test-retest reliability of .70 for the correct reasoning scale and of .75 for the misconception scale (Garfield, 2003) makes it a good choice for research (Nunnally & Bernstein, 1994).

Another important strength of the SRA is that it places significant focus on assessing one's understanding of the statistical concepts rather than just the application of calculations. In fact, no calculations are necessary. The entire instrument is in a multiple-choice format, which makes it a good instrument of choice both for classroom assessments and for research. In this multiple-choice format, the true answers were embedded amongst incorrect answers (foils) whose content was based on erroneous but plausible answers given by actual students in an early round of the instrument's development.

Variables affecting statistical competence

Several researchers have used the SRA to assess statistical competence in a wide range of populations, and several important findings have emerged. Critically for the present purposes, Liu's research (1998), as reported in Garfield (2003), has demonstrated a clear gender effect, where males outperform females in their ability to avoid misconceptions. The effect was marginally significant for correct reasoning. Also using the SRA, Tempeelar et al. (2006) replicated the gender effect in statistics both for the ability to reason correctly ($p < .001$, $d = .24$) and the ability to avoid well-known misconceptions ($p < .001$, $d = .27$). These differences were found despite little or no difference in prior education, and despite the fact that all participants were taking their first course in statistics. It is thus not clear what causes this gender gap. As Tempelaar et al. (2006) note, what is especially puzzling is the fact that this gender difference occurs despite similar educational backgrounds of the males and females. However, potential factors of interest such as individual differences in cognitive ability and motivation were not taken into account in their research. Nonetheless, similar results have been found in mathematics. Specifically, Byrnes and Takahira (1993) reported that, even when obtaining the same grades in the classroom, females nonetheless performed more poorly than men on the quantitative section of the SAT.

Although background education does not explain the gender gap, many researchers have examined the impact of specific training and general class experience on statistical reasoning. In four experiments, Fong, Krantz, and Nisbett (1986) examined the extent to which people use the law of large numbers in everyday problems, and whether the frequency and the quality of their statistical reasoning can be improved through specific short-term training (Experiments 1 and 2) and through formal in-class experience (Experiments 3 and 4). In their experiments, participants read three different types of scenarios: probabilistic (e.g., lottery, where randomness is obvious), objective (e.g., sports achievement, car reliability), and subjective (e.g., what college course to take), and were asked to explain the outcomes. Participants' tendency to explain the scenarios – such as why a meal may not be as extraordinary on a second visit to a restaurant – in statistical terms (rather than blaming the chef!) improved greatly with specific short-term training sessions on the law of large numbers as well as with additional course experience. For instance, where novices rarely used statistical terms to explain the scenarios, those having completed at least one course in statistics provided explanations rooted in statistical terms – such as "regression to the mean" – up to 40% of the time, while those at the doctoral level provided statistical explanations closer to 80% of the time. However, Nisbett, Krantz, Jepson, and Kunda (1983) also warned that it might be the experience in a domain rather than the level of experience in statistics that encourages people to look at a problem in a statistical rather than in a deterministic fashion. Furthermore, the same reasoning skills do not develop in every domain (Gray & Mill, 1990; Nisbett, Fong, Lehman, & Cheng, 1987). Fong et al.'s (1986) findings also fail to control for cognitive ability. With regard to gender, their research sheds no light on that issue. Unfortunately, gender of participants was either not reported (Experiments 1-3) or limited to males (Experiment 4).

Quilici and Mayer (1996) also relied on the specific short-term training of a group of participants who had taken zero or one course in statistics. As part of the training, they had participants study examples of t-test, correlation and chi-square problems that either emphasized the structure of the problems (e.g., all correlation examples grouped together on the same page) or the surface features of the problems (e.g., all problems related to the weather presented on the same page) prior to completing a sorting task in which participants were to place each of 12 statistical problems into groups with the other problems they best went with. The sorting task allows inferences to be made about how participants are thinking about these problems, and how they are representing them in memory. Although surface features (e.g., weather in the example above) are more salient, it is the ability to recognize structurally related problems that is the important skill in mathematical problem solving (Polya, 1945). Quilici and Mayer (1996) were the first to extend this task to statistical word problems, demonstrating that appropriate training lead participants to sort statistical word problems based on their deep structure rather than based on their surface similarity. Indeed, not only did training with structure-emphasizing examples lead participants to categorize the problems based on their structural features more often, it also lead to greater application of the appropriate statistical test for those in the structure-emphasizing group. Their findings were qualified by the fact that training was much more beneficial for lower ability students than for higher ability students. Unfortunately, gender was not included as an independent variable. Also, for those interested in statistical literacy in general, it is noteworthy that those training sessions were highly specific, covering only the notion of the law of large numbers (Fong et al, 1986) or a few targeted inferential tests (Quilici & Mayer, 1996). This narrow focus could be the reason behind the finding of a training effect.

Although knowing about the performance of participants on a narrow statistical task may be interesting at the experimental level, the findings cannot be generalized easily and do not reflect the breadth of knowledge necessary to be considered statistically literate in today's society. In contrast to the dependent variables used in the studies above, the SRA addresses multiple areas of statistical reasoning, including the law of large numbers, amongst others (e.g., averages, probabilities, correlation versus causation, etc.). Of course, such breadth of knowledge cannot be communicated in a single training session. It is thus realistic to assume that considerably more training is necessary to generate a significant improvement. This hypothesis, along with others, will be tested in the current series of experiments. That being said, the effect of targeted training may still only have a limited effect as it is well known that even experts fail to achieve a perfect score on some statistical tasks (e.g., Hoffrage et al., 2002; Kahnemann, Tversky, & Slovic, 1982). Again, formal education does not appear sufficient on its own to ensure proper use outside the classroom. Yet, we know that "effective transfer is critical here because statistical reasoning is applicable across a wide variety of domains and in daily life; statistical reasoning skill is of little value if it can only be applied in the statistics classroom" (Lovett, 2001, p. 347).

Overarching Research Goals

Given the importance of statistical competence in today's world, and knowing that any discrepancy may impact the long-term success of any group of lower ability, my research aimed to provide a strong test of whether gender differences exist in this domain. To understand the challenges for statistical education, as well as the changes and actions that may be required, I proposed an examination of factors that may help explain and close the gender gap in statistics. With the goal of identifying more precisely where the gap lies, the impact of experience in statistics was examined, especially as an interaction of experience with gender would impact how one might go about redressing it. Individual differences, stereotype threat, and task format were also considered. To accomplish this, six studies were conducted. In Study 1, it was hypothesized that experience and individual differences within each gender group would help explain, at least in part, the gender gap. Participants completed the SRA as well as multiple measures of thinking dispositions and cognitive ability. Foreshadowing the results, since gender was found to have an effect on statistical reasoning above and beyond training and individual differences, the next set of studies focused on the potential negative impact of stereotype threat, a phenomenon that has been demonstrated to decrease performance of females in mathematics and other science, technology, engineering, and mathematics (STEM) disciplines (Shapiro & Williams, 2012). In Studies 2, 3 and 4, the self-affirmation technique promoted by Martens et al (2006) in mathematics and spatial reasoning and by Miyake et al (2010) in physics as an effective tool to counter the effect of stereotype threat was tested, in different settings and using different statistical assessment tools. Finally, studies 5 and 6 explored how participants mentally represent statistical problems in memory, and tested the effect of using general versus constrained instructions on performance. As each of these lines of work necessitated the examination of different literatures, the reviews for each section will be provided separately.

STUDY 1

Using the SRA, both Garfield (2003) and Tempelaar et al (2006) have demonstrated a disturbing gender gap in statistics. However, each of their research has focused on a different point in time. Garfield (2003) focused more specifically on testing knowledge at the end of an introductory course in statistics. When administering the SRA at the end of the semester, she found that males performed better than females. In an attempt to rule out the role of instruction and to clearly tap into any misconceptions that students might hold, Tempelaar et al. (2006) administered the SRA at the beginning of a semester. Despite recruiting participants with homogeneous education background, the gender gap was found once more. Also, a weak negative correlation was found between the SRA and effort-based measures (i.e., homework), and a weak positive correlation was found between the correct reasoning score on the SRA and the final exam. In contrast, Garfield (2003) had found no correlation between performance on the SRA and course performance. However, by limiting the range of experience in their sample, the question of knowing if the gender gap is persistent or transient remains unanswered.

For other research focusing specifically on the role of experience in improving the quality of statistical reasoning (e.g., Fong, Krantz, & Nisbett, 1986; Quilici & Mayer, 1996, 2002, Hogan & Rabinowitz, 2008, 2009; Lavigne, Salkind, & Yan, 2008), it is then gender that is not included nor reported as a variable of interest. Thus, it is not possible to know whether the two factors of gender and experience interact. This study aims to shed light on this question. Also, instead of focusing on short and pointed training sessions, such as used in Fong, Krantz, and Nisbett (1986) and Quilici & Mayer (1996), the more ecologically valid approach of taking current experience in statistics as predictor variable, as utilized by Hogan and Rabinowitz (2008, 2009), Quilici and Mayer (2002) and Nisbett et al. (1983), was used.

Furthermore, a question that has not been addressed in previous studies is whether the gender gap will decrease (or increase) with additional experience in statistics. The first study in this dissertation aimed to shed light on this question. If a gender gap exists, will further training advantage one gender in the process, or will the difference remain constant? To explore this question, males and females with a range of experience, operationally defined as the number of statistics courses taken in university, completed the SRA. Both a gender gap and a beneficial impact of training were expected. However, it was unknown whether an interaction would be found. On one hand, it could be expected that only those females with greater ability in statistics will actually go through the process of taking more than one course in statistics, which may technically lead to a reduction in the performance gap. On the other hand, it is possible that the gender gap will simply continue. These questions were addressed in the first study.

A second goal of this first study was to better understand the role of individual differences in statistical literacy, which has become a dominant theme in cognitive psychology in general, and in reasoning research in particular. Here, many prominent reasoning theorists argue that the product of reasoning performance is the sum of more than just simple abilities (e.g., Stanovich, 2001; Baron, 1985; Ennis, 1987). Indeed, Stanovich and West's research, with their colleagues (1997, 1998, 1999, & 2007) has demonstrated that reasoning outcomes are not fully explained by cognitive abilities alone. They find that after controlling for cognitive ability, a substantial portion of the remaining variance can be explained by thinking dispositions – which can be described simply as intellectual inclinations that benefit good, productive thinking (Ritchhart, 2001). In support of the focus of this dissertation, Hawkins (1997) highlights the relevance of thinking dispositions for statistical reasoning:

> "As statisticians, we are aware that the media, our policymakers, members of the general public, our students, and even ourselves on occasions, are prey to many statistical and probabilistic misconceptions. Some of these misconceptions seem to be reasonably easy to address. Research shows, however, that others remain deep-seated and resistant to change. In fact, it is not only peoples' misconceptions that we need to worry about. To be statistically literate, a person must have not only reliable understanding, *but also an inclination for using that understanding in everyday reasoning*." [emphasis added]

Thinking dispositions are an attractive focus of research because they are seen as more malleable than cognitive abilities (Stanovich, 2001; Baron, 1985) and as holding the power to regulate the use of cognitive abilities to their full potential (e.g., Cacioppo, Petty, & Kao, 1984; Stanovich, 2009). If that hypothesis holds true, this signifies that people's performance on a reasoning test can be improved simply by influencing their level of motivation and dedication to the task. Alternatively, if two individuals possess the same amount of cognitive abilities, the one with the highest dispositions toward the reasoning task should perform better. This explains the importance Stanovich gives to thinking dispositions in his most recent model of reasoning (2009). Indeed, Stanovich's model states that the level of thinking dispositions indirectly influences reasoning performance by directly regulating the display of cognitive abilities.

However, as far as I am aware, the appropriateness of this model for statistical reasoning has not been tested directly. Thus, a secondary goal of this preliminary study was to examine the role of thinking dispositions and cognitive abilities in statistical reasoning using a confirmatory structural equation modeling approach. If the model holds for statistical reasoning, the moderating influence of gender on the interplay between cognitive ability and thinking dispositions will be examined. Finally, to eliminate the possibility that women did not engage fully in the task because they misjudged their performance to be good, confidence ratings were obtained after each question. A good awareness of their performance would result in a high correlation between their performance and confidence levels.

Method

Participants

Two hundred and one University of Waterloo undergraduate and graduate students proficient in English participated for course credit or monetary remuneration. Following Frederick (2005), two participants with scores below 10 on the Wonderlic Personnel Test were eliminated from the analysis, reducing the sample to 199 participants (92 males, 107 females; M_{age} = 21.57, *SD* = 4.59). Participants had varying levels of experience in statistics, as measured by the number of statistics courses they had previously taken (0: N = 76, 1: N = 46, 2 or more: N = 77), and – based on participants' responses to a questionnaire item (see Materials below) – came from fields deemed generally non-quantitative (e.g., child care, music, art, English, philosophy) to extremely quantitative (e.g., mathematics, statistics) [Generally non-quantitative = 10 (5%), Minimally quantitative = 22 (11%), Moderately quantitative = 59 (30%), Highly quantitative = 82 (41%), Extremely quantitative = 26 (13%)].

Design and Materials

To examine the influence of gender and experience on statistical reasoning, a 2 (gender) x 3 (experience) between-subjects design was used to analyse performance and confidence. Confidence ratings were collected to gauge performance awareness and to assess calibration (i.e., being more confident when correct and being less confident when incorrect). Performance on the statistical task was further analysed in light of thinking dispositions and cognitive abilities.

Statistical Task. The Statistical Reasoning Assessment (SRA: Garfield, 2003) was used as the main task. This test comprises 20 word problems assessing various components of statistical reasoning, such as choosing an appropriate average, understanding sampling variability, and distinguishing between correlation and causation. All answers are given using a multiple-choice format. For each question, some of the choices represent correct reasoning, while other choices represent some prevalent misconceptions (i.e., beliefs, interpretation or understanding that are not only mistaken, but also resistant to change; Chi & Roscoe, 2002; Fischbein, 1987). Examples of misconceptions in statistics include thinking that groups cannot be compared if they are not the same size, failing to take outliers into consideration when computing the mean, judging probabilities based on representativeness, and assuming that small samples are as good as large ones for drawing conclusions. The presence of correct, incorrect and misconception-related items in the set of answer choices allows the calculation of two scores: a "correct reasoning" score (CR) and a "misconception" score (MISC). Each score is a weighted average of performance on eight components for each scale (see Garfield, 2003, and Tempelaar et al, 2006, for more details on scoring). A copy of the test and its sixteen subscales is available in Appendix A.

Performance awareness and calibration were also assessed. To do so, participants were prompted to rate their confidence in the accuracy of their answer after each question, indicating their rating on a 6-point scale (ranging from 1 = not confident at all to 6 = very confident). An overall confidence score was obtained for each participant by averaging the ratings across all 20 questions.

Individual Differences. Measures of thinking dispositions (i.e., intellectual inclinations that benefit good, productive thinking: Ritchhart, 2001), and measures of cognitive ability were used. The following thinking dispositions scales were used: the Preference for Numerical Information Scale (PNI: Viswanathan, 1993; Coefficient alpha reported by the creator of the scale = .94), the Need for Cognition Scale (NC: Cacioppo, Petty, & Kao, 1984; Coefficient alpha reported by the creators of the scale = .90), and the Actively Open-Minded Thinking Scale

(AOT: Stanovich and West, 1997, 1998, 2007; Sà, West, & Stanovich, 1999; Coefficient alpha reported by the creators of the scale = ranging from .81 to .88). To measure verbal, numerical and general cognitive abilities, the Vocabulary Checklist-with-Foils task (VOC: used as a proxy for cognitive ability in Stanovich & West, 1997; split-half reliability reported by the previous authors = .87), the Numeracy Scale (NUM: Lipkus, Samsa, & Rimer, 2001; Coefficient alpha reported by the creators of the scale = ranging from 0.70 to 0.75), the Cognitive Reflection Test (CRT: Frederick, 2005; no psychometric information available in the literature), and the Wonderlic Personnel Test – Form A (WPT: Wonderlic Inc., 2002; Coefficient alpha reported in the user's manual = ranging from .88 to .94) were used[1]. All scales (except for the WPT due to copyright limitations) are available in Appendix B.

Demographic information was also collected, including some questions drawn from Schield (2005). The information of interest included gender, age, university level, number of statistics courses completed, and number of research method courses completed. Participants were also asked to self-report their level of comfort with formal statistics and with informal statistics on a 4-point scale ranging from very uncomfortable to very comfortable, as well as the level of quantitative knowledge required in their field, reporting this value on a 5-point scale ranging from 'generally non-quantitative' (e.g., child care, music, art, English, philosophy) to 'moderately quantitative' (e.g., psychology, sociology, market research, forecasting) to 'extremely quantitative' (e.g., mathematics, statistics). This questionnaire is also available in Appendix B.

[1] As the use of those measures as covariates in the analyses did not change any of the patterns of findings in Studies 1, 2 and 4, the results of those analyses are not reported in the main results sections. Instead, for each of the six

Procedures

The study was conducted in two parts. The first part occurred online, at the participant's convenience prior to coming to the lab, and was scheduled for 30 minutes. Participants filled out three self-report questionnaires: PNI, NC, and AOT, as well as demographic information. The second part of the study occurred in lab and was scheduled for 60 minutes. Five paper-pencil tasks were completed in this order: 1) SRA (Garfield, 2003), along with confidence ratings, 2) VOC, 3) CRT, 4) NUM and 5) WPT – Form A. Consent was obtained from each participant at the start of each portion of the study, and feedback was given after the in-lab session was completed.

Results

To test the stated hypotheses, multiple analyses were necessary. The initial set of analyses tested the role of experience and gender on statistical reasoning. As a first step, the performance data from the SRA were analysed. As noted above, the SRA allows the computation of two separate subscales: a correct reasoning (CR) score, and a misconception (MISC) score. These two scores were analysed using a 2 (gender) x 3 (experience) between-subjects ANOVA. As a second step, the same analyses were repeated on confidence ratings. All descriptive statistics are available in Table 1.

The subsequent set of analyses was concerned with the relations among cognitive ability, thinking dispositions, and statistical reasoning. Firstly, zero-order correlations were obtained. Secondly, the appropriateness of Stanovich's (2009) tri-partite model was tested using confirmatory structural equation modelling. As appropriate, the role of gender as a predictor was examined.

Table 1

Descriptive Statistics – Study 1 – Performance & Confidence

		Male			**Female**			**Total**		
	# Stats courses taken	n	*Mean*	*SD*	n	*Mean*	*SD*	n	*Mean*	*SD*
Correct reasoning Scale	0	**30**	.57	.18	**46**	.49	.16	**76**	.52	.17
	1	**17**	.67	.11	**29**	.55	.12	**46**	.59	.13
	2^+	**45**	.73	.13	**32**	.61	.13	**77**	.67	.14
	Total	**92**	**.66**	**.16**	**107**	**.54**	**.15**	**199**	**.60**	**.16**
Misconception Scale	0		.29	.12		.32	.12		.31	.12
	1		.21	.12		.31	.10		.27	.12
	2^+		.18	.09		.26	.09		.22	.10
	Total		**.23**	**.12**		**.30**	**.11**		**.27**	**.12**
Confidence	0		4.72	.51		4.44	.85		4.55	.75
	1		5.21	.49		4.77	.50		4.93	.54
	2^+		5.25	.47		4.83	.63		5.08	.58
	Total		**5.07**	**.54**		**4.65**	**.72**		**4.84**	**.68**

Effect of Gender and Training on Performance

As previous studies using the SRA found a gender difference (e.g., Teempelaar et al., 2006; Garfield, 2003), we expected that males would perform better than females on the SRA. However, as those studies focused uniquely on the first course in statistics, and given that training and experience have been shown to improve statistical reasoning (e.g., Fong, Krantz, & Nisbett, 1986; Quilici, & Mayer, 1996, 2002; Rabinowitz & Hogan, 2008,2009), we predicted that the gender gap could vary with increased experience. The data were analysed with a 2 (gender) x 3 (experience) analysis of variance, both for the correct reasoning score (CR) and for the misconceptions score (MISC).

Overall, males performed better than females (see Table 1), scoring higher on the CR scale by an average of 12% across experience levels (M_m = 0.66, *SD* = 0.16; M_f = 0.54, *SD* = 0.15), $F(1, 193) = 26.31$ $MSE = .020$, $p < .001$, $\eta^2_p = .120$, and committing fewer mistakes, thus scoring lower on the MISC scale (M_m = 0.23, *SD* = 0.12; M_f= 0.30, *SD* = 0.11) by an average of 7% across experience levels, $F(1, 193) = 16.73$, $MSE = .012$, $p < .001$, $\eta^2_p = .080$, which is consistent with our hypothesis. This main effect of gender occurred while effectively controlling for experience in the comparison above and, even after controlling for intelligence using WPT as the covariate, remained statistically significant for correct reasoning ($p < .001$, $\eta^2_p = .227$), but not for misconceptions ($p = .22$, $\eta^2_p = .008$).

As predicted, increased experience was associated with better performance. Indeed, each extra course in statistics was associated with improved correct reasoning, $F(2, 193) = 16.41$, $MSE = .020$, $p < .001$, $\eta^2_p = .145$, which was confirmed with multiple comparisons, revealing that each additional level of experience corresponded to significantly higher performance than the previous level (*Tukey HSD*, p < .05). Misconceptions also varied significantly with increased experience, $F(2, 193) = 9.87$, $MSE = .012$, $p < .001$, $\eta^2_p = .093$. Specifically, misconceptions were significantly lower with increased experience, but only for those having taken at least two courses in statistics (*Tukey HSD*, $p < .05$). Indeed, those with one course in statistics did not fare any better than those with no experience in statistics (*Tukey HSD*, p = .17). Thus, this suggests that misconceptions may require more experience to change than correct reasoning. This finding is consistent with the literature on conceptual change (Chi & Roscoe, 2002), which has shown that misconceptions can be highly resistant to change. It is also worth noting that performance on neither of the subscales came close to ceiling (CR) or floor (MISC) with additional experience. Importantly, the gender gap did not decrease with experience, as no interaction was found with either correct reasoning, $F(2, 193) = .28$, $MSE = .020$, $p =.757$, $\eta^2_p = .003$, or misconceptions, $F(2, 193) = 1.44$, $MSE = .012$, $p =.241$, $\eta^2_p = .015$.

Effect of Gender and Training on Confidence

If participants are well calibrated, i.e., if their confidence is an accurate reflection of their performance (e.g., low confidence when answer is incorrect, high confidence when answer is correct), then the same pattern of findings should be present in the analysis of variance of the confidence ratings, and the correlation between performance and confidence should approach 1.

A 2 (gender) x 3 (experience) ANOVA revealed the same overall pattern as found with the performance data, with two significant main effects and no interaction. Reflecting performance, males ($M = 5.07$, $SD = .539$) were more confident than females ($M = 4.65$, $SD= .724$), $F(1, 193) = 16.91$, $MSE = .379$, $p < .001$, $\eta^2_p = .081$, and increased experience led to greater confidence ($M_0 = 4.55$, $SD = .746$; $M_1 = 4.93$, $SD = .538$; $M_{2+} = 5.08$, $SD = .575$), $F(2, 193) = 11.57$, $MSE = .379$, $p < .001$, $\eta^2_p = .107$. Nonetheless, closer examination of the effect of experience revealed a different pattern. Whereas experience continued to have incremental effects on performance with each statistics course taken, confidence increased significantly after having taken one course in statistics (*Tukey HSD*, $p < .01$) and then levelled off, as no further difference was found with increasing experience (*Tukey HSD*, $p = .42$). At this point, we cannot differentiate between the possibilities of those having taken one course in statistics being overconfident versus those having taken three courses in statistics being under-confident, although a preference is given to the former possibility due to past research demonstrating people's bias toward overconfidence (e.g., Fischhoff, Slovic, & Lichtenstein, 1977; Lichtenstein

& Fischhoff, 1977). It is also interesting to note that males' confidence was not as strongly correlated with their performance (*r* = .24, *p* = .023) as females' confidence was with their performance (*r* = .48, *p* < .001); *z* = 1.91, *p* = .056. A scatterplot summarizing the performance / confidence results are presented in Figure 1.

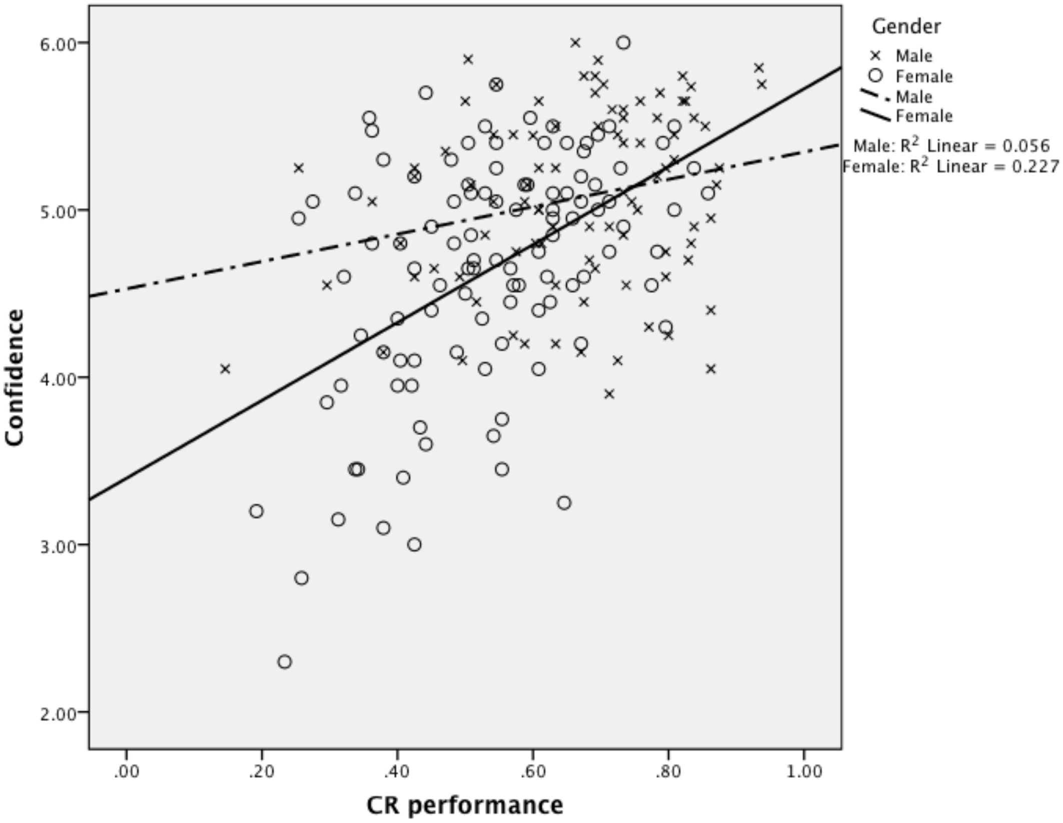


Figure 1. Scatterplot of the association between performance and confidence, by gender. Study 1.

Individual Differences - Correlations

A first examination of the correlation matrix (see Table 2) revealed that all associations are in the predicted direction, with all measures of individual differences (except VOC) correlating positively with correct reasoning and confidence, and negatively with misconceptions. For correct reasoning, performance correlated between *r* = .20 (AOT) and *r* =

.38 (PNI) with thinking dispositions, while correlating between *r* = .14 (VOC) and *r* = .56 (CRT) with cognitive ability[2]. The correlation with the Vocabulary task was exceptionally low in comparison to the correlations with the CRT and the WPT (*r* = .55). This is particularly surprising, as Stanovich and West (1997) have used this Vocabulary task as a proxy for cognitive ability without any other measures to check their assumption. For misconceptions, the correlations were negative, as they should be, ranging from *r* = -.14 (AOT) to *r* = -.25 (NC) for thinking dispositions, while correlating from *r* = -.14 (VOC) to *r* = -.41 (CRT) for cognitive ability. Finally, for confidence, the correlations ranged from *r* = .08 (AOT) to *r* = .48 (PNI) for thinking dispositions, while correlating from *r* = -.01 (VOC) to *r* = .47 (CRT) for cognitive ability. Overall, if the correlations from the VOC are disregarded, cognitive abilities are more highly associated with correct reasoning; the CRT is the best predictor of misconceptions use; and high scores on the PNI and on the CRT are the most predictive of a high level of confidence.

[2] Some readers may raise an eyebrow at the view of CRT being considered as a measure of cognitive ability rather than as a measure of thinking dispositions. However, Frederick (2005) was agnostic in its categorization of his measure as being one or the other: "I have proposed that the CRT measures 'cognitive reflection'—the *ability or disposition* to resist reporting the response that first comes to mind" [emphasis added] (p.35)." When testing whether CRT should be considered an indicator of CA, of TD, or of both, an additional factor loading was added from TD to CRT. The obtained parameter estimates clearly support the view of using the CRT solely as an indicator of CA. Specifically, whereas the factor loading between CA and CRT (.76) was significant (p < .001), the loading between TD and CRT stood merely at .03 and was clearly non-significant. On that basis, the factor loading between TD and CRT was considered disconfirmed, and the possibility of treating it as an indicator of both latent variables was abandoned.

Table 2

Correlation Matrix – Study 1

Subscale	**Mean (SD)**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1. CR	.60 (.16)	--	-.56**	.45**	.38**	.26**	.20**	.55**	.56**	.44**	.14*
2. MISC	.27 (.12)		--	-.27**	-.24**	-.25**	-.14	-.22**	-.41**	-.22**	-.14*
3. Conf	4.84 (.68)			--	.48**	.30**	.08	.33**	.47**	.37**	-.01
4. PNI	84.67 (15.66)				--	.44**	.19**	.42**	.43**	.43**	-.06
5. NC	73.31 (12.58)					--	24**	.27**	.37**	.30**	.15*
6. AOT	146.13 (8.13)						--	.22**	.15*	.12	.01
7. WPT	28.41 (6.07)							--	.59**	.52**	.15*
8. CRT	1.56 (1.17)								--	.53**	.06
9. NUM	9.78 (1.63)									--	.02
10. VOC	21.12 (4.95)										--

Note. * *p* < .05. ** *p* < .01.

Structural Equation Model – Gender and Individual Differences

The persistence of the gender gap despite increased training is an alarming finding. Why is this occurring? What role do individual differences in thinking dispositions and cognitive ability play in statistical reasoning? According to Stanovich's (2009) tri-partite model of reasoning, beyond the expected positive impact of higher cognitive ability on the quality of reasoning, higher thinking dispositions also affect reasoning indirectly by influencing the use one makes of their own cognitive ability. In the current context, the question of interest is how gender influences this process and the final reasoning performance.

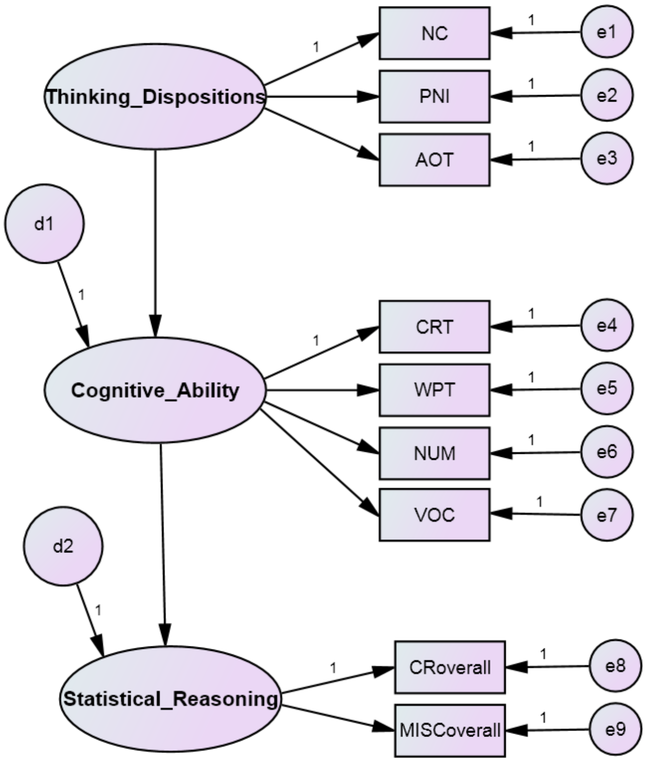


Figure 2. First version of the Structural Equation Model based on Stanovich (2009).

To examine the relations between thinking dispositions, cognitive ability and statistical reasoning, a structural equation model (presented in Figure 2) was used. Structural equation models are composite models that include both a measurement model and a path model. The measurement model illustrates the relation between the latent variables (unmeasured) and their specific indicators (measured). For instance, in the current study, cognitive ability was captured through four measured indicators: the Wonderlic Personnel Test, the Numeracy Scale, the Cognitive Reflection Test, and the Vocabulary-Test-with-Foils. Similarly, thinking dispositions were captured through three measured indicators: the Preference for Numerical Information, the Need for Cognition Scale, and the Actively Open-Minded Scale.

The path model illustrates the relations among the main constructs of interest. In this case, the path model includes only latent variables and depicts the causal model proposed by Stanovich (2009), which states that thinking dispositions affect the quality of our reasoning indirectly through the influence they exert on the deployment and use of available cognitive abilities. In addition to its flexibility, the main advantage of using a structural equation model rests on the fact that relations among the latent variables are corrected for measurement error, which is not true when using regression analyses (Kline, 2011).

The first step in the use of this structural equation model was to test the generalizability of Stanovich's (2009) tri-partite model to the area of statistical reasoning. Support for the tri-partite model would come from finding that the proposed model fits the data well. Fit indices are calculated based on how closely the model allows the reproduction of the correlations present in the actual data. The closer the reproduced correlations are to the actual data, the better the fit. Next, if the fit of the general model were acceptable, the equivalence of the reasoning process across gender would be ascertained. To do so, path coefficients are set to be equal across gender. If the fit remains good, this suggests that the pattern of relations is equivalent across genders. However, if the fit becomes poor, this suggests that the genders have different patterns among the latent variables. Finally, if the process can be shown to be equivalent, the influence of gender on each of the three parts of the model (i.e., TD, CA, SR) can be examined by including gender as a measured exogenous categorical predictor in the model.

Testing the appropriateness of the model for statistical reasoning. In their work to substantiate the role of thinking dispositions in reasoning, Stanovich and his colleagues (e.g., Stanovich & West, 1997, 1998b; Toplak & Stanovich, 2002, 2003) have relied on multiple regression analysis. Their main argument to support the role of thinking dispositions is that a significant portion of the variance left unexplained by cognitive ability can always be explained by thinking dispositions. However, a main limitation of the regression approach is that its results do not correct for measurement errors (Kline, 2011). In contrast, structural equation models explicitly depict the difference between constructs that are latent and indicators that are observed. By definition, we know that the measures used as indicators are an imperfect snapshot of those constructs. SEM takes those measurement errors into consideration, correcting the resulting path coefficients between the latent constructs for attenuation. Also, each measure is given a different weight to represent its quality in relation to the construct. In this sense, structural equation modelling is a more rigorous method of analysis (Bollen, 1989; Bullock, Harlow, & Mulaik, 1994; Jöreskog and Sörbom, 1989).

In Stanovich's (2009) tri-partite model of reasoning, one important assumption is that thinking dispositions influence the expression of cognitive ability, which in turn determines reasoning performance. In fact, this model assumes no direct path between thinking dispositions and statistical reasoning. This path model (see Figure 2), complemented by the aforementioned indicators, is the basis for the confirmatory test of the proposed model of reasoning.

The first model (see Figure 2) included all indicators for each latent variable. Despite a significant *Chi-square* ($\chi^2 = 46.32$, $df = 25$, $p = .006$), which often occurs as the sample size increases, the other fit indices reveal a satisfactory fit. The *comparative fit index* is above .95 (*CFI* = .955). The *root mean square error of approximation* is below .08 (*RMSEA* = .066) and the related *p of close fit* – which indicates whether the difference of the obtained *RMSEA* value from close fit is attributable to sampling error – is above .05 (*pclose* = .179). All estimates (except VOC) are significant (p's < .001), supporting the appropriateness of this dual-process model to the area of statistical reasoning. However, one of the indicators has a non-significant factor loading. The regression weight for VOC is only .12 (p > .05), which indicates that it is not an appropriate indicator of cognitive ability in the current model. For this reason, this indicator was removed and the model was re-estimated.

For this second model (see Figure 3), the obtained *Chi-square* value is non-significant ($\chi^2 = 28.503$, $df = 18$, $p = .055$), which is a very good indication of the fit of the model. Of course, the other fit indices concur on this finding of good fit (*CFI* = .977; *RMSEA* = .054, *pclose* = .389). Another sign of the usefulness of removing VOC from the list of indicators is the fact that the *expected cross-validation index* (ECVI), a fit index that takes parsimony into account, dropped noticeably from the first to the second model (.436 to .326). Overall, this model explains 50% of the variance in statistical reasoning as measured by the SRA in this sample. Given the significant paths between TD and CA, as well as between CA and SR, this analysis lends support to Stanovich's idea that thinking dispositions regulate the manifestation of the algorithmic level represented by cognitive ability. However, one possible alternative is worth testing.

The obvious alternative model is that thinking dispositions may have a direct effect on statistical reasoning. To test this possibility, a path was added between TD and SR in the model above. The addition of that path does not alter the fit dramatically ($\chi^2 = 28.232$, $df = 17$, $p = .042$; *CFI* = .976; *RMSEA* = .058, *pclose* = .333; *ECVI* = .335). Importantly, the added path, estimated to be .08, does not reach significance. Thus, despite the possibility that a small direct effect may exist between thinking dispositions and statistical reasoning, the assumption of the absence of a direct effect between TD and SR is sufficiently supported to continue omitting it.

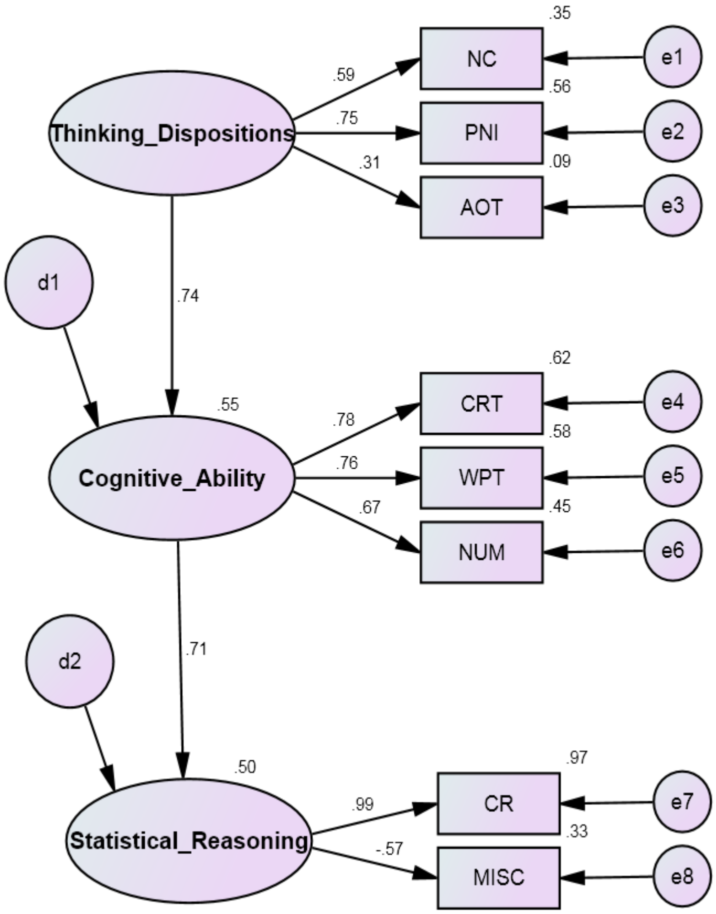


Figure 3. Second version of the Structural Equation Model based on Stanovich (2009).

Process equivalence. To ensure that the same reasoning process applies both to males and females, a multi-group SEM analysis (Arbuckle, 2009) was also used. In this model, data is analysed concurrently for each gender, with the particularity that the critical paths (i.e., the path between TD and CA, and the path between CA and SR) are set to be equal across genders. If the equivalence assumptions added are not viable, the fit indices will indicate poor fit. In contrast, all fit indices remained good (χ^2 = 43.82, *df* = 38, *p* = .238; *CFI* = .986; *RMSEA* = .028; *pclose* = .857; *ECVI* = .730), indicating that the model proposed by Stanovich is applicable to both genders.

Gender influence. The remaining question regards how gender exerts influence on this reasoning process. To test the total effect of gender on statistical reasoning, the original model was thus modified to include this observed categorical predictor variable, with males coded as 0, and females coded as 1 (see Figure 4). The analysis revealed gender as influencing statistical reasoning in multiple ways in this well-fitting model (χ^2 = 42.67, *df* = 23, *p* = .008; *CFI* = .961; *RMSEA* = .066, *pclose* = .186; *ECVI* = .438). First, being female has a significant negative impact on thinking dispositions (-.33, *p* <.001), on cognitive abilities (-.15, *p* = .065), and on statistical reasoning (-.16, *p* = .016). Combining this information with the significant paths between TD, CA and SR, being female had a negative impact on SR in three separate ways. First, the lower thinking dispositions of females decreased the use of cognitive ability to properly solve the statistical problems [indirect path = (-.33)(.69)(.67) = -.15]. Second, even when holding thinking dispositions constant, there was a further effect of gender on cognitive abilities, which also predicted lower performance in statistical reasoning [indirect path = (-.15)(.67) = -.10]. Finally, even when controlling for cognitive ability, gender had a direct effect (-.16) on statistical reasoning that cannot be explained by differences in cognitive ability, or differences in thinking dispositions. That is, of the total effect (-.41) of gender on statistical reasoning, -.15 (37%) is attributable to thinking dispositions, -.10 (24%) is attributable to cognitive ability (excluding its role as a mediator of the effect of cognitive dispositions), and -.16 (39%) remains that is not explained by these two variables.

Taken together, these results indicate that multiple approaches can be used to attempt to raise the performance of females in statistics. Based on the current results, those approaches could include interventions to raise thinking dispositions, interventions to improve cognitive ability, and other interventions that may have a direct influence on statistical reasoning. However, given that multiple routes have the potential to benefit statistical reasoning performance, any attempt to influence statistical reasoning indirectly or directly will ever only address approximately one-third of the overall effect, as shown above by the proportion of the total effect attributed to each of the three effects.

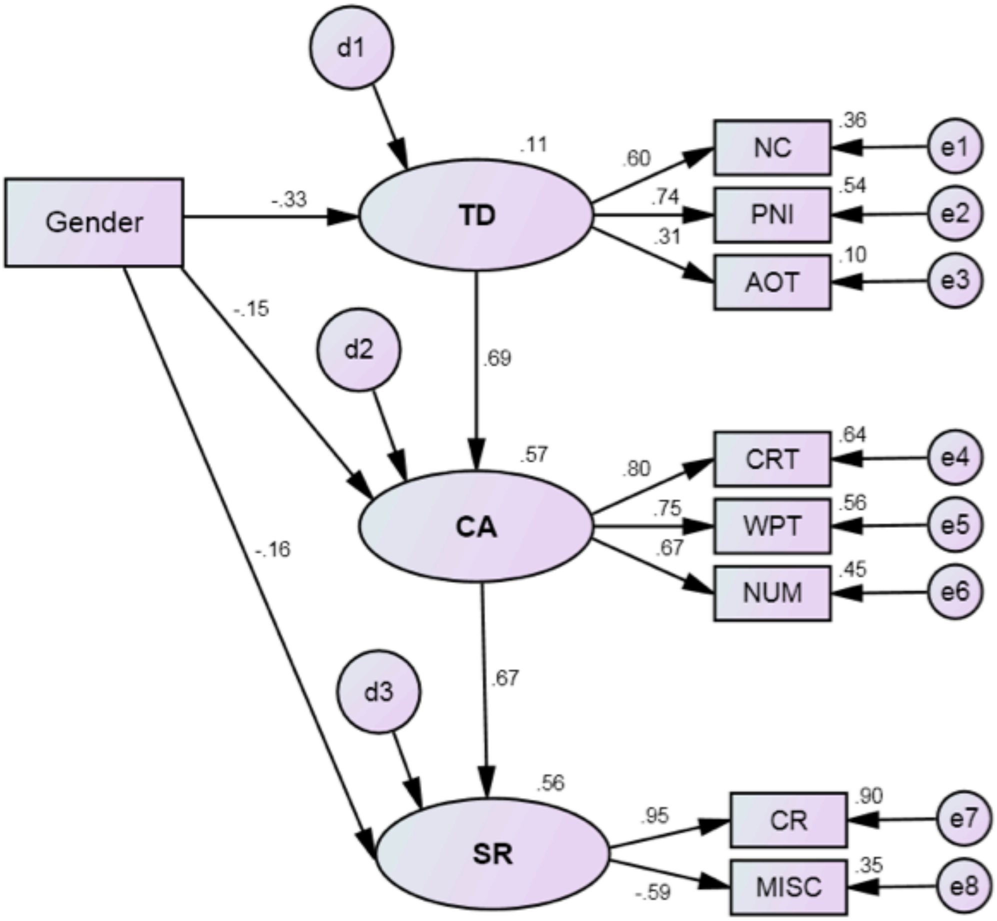


Figure 4. Examining the influence of gender on statistical reasoning.

Discussion

In this first study, by controlling for experience and individual differences, I provided the strongest evidence to date for the existence of a persistent gender gap in statistics. Even though increased experience in statistics was associated with an increase in performance overall, it was not sufficient to close the gender gap. For instance, only women having taken two courses in statistics reached the level of performance of men with no experience in statistics. At the same level of experience, men surpassed them easily, both in their ability to display correct statistical reasoning and in their ability to avoid misconceptions. Of course, the cross-sectional nature of the sample limits the conclusions that can be drawn about the role of experience, as it is possible that a self-selection bias may have influenced the composition of the groups at each level of experience. For instance, it is possible that only those higher in cognitive ability keep taking statistics beyond the mandatory introductory class. However, it is useful to note that the difference in performance across genders remained significant even after controlling for cognitive ability. Also notable was how much room for improvement was left for both genders, even after completing two courses in statistics. This is consistent with prior research by Fong, Krantz and Nisbett (1986). In their study, participants with 1 to 3 courses in statistics referred to statistical concepts such as regression to the mean and law of large numbers to explain diverse scenarios involving variation – one of the most important ideas in statistics – no more than 40% of the time. Even those at the doctoral level used statistically grounded rather than deterministic explanations no more than 80% of the time. In their study, Fong et al. did not examine the role of gender, however.

Gender was also prominent in the examination of patterns of confidence. Whereas women's level of confidence was generally consistent with their level of performance ($r = .48$), men’s confidence was generally high regardless of their performance (r = .24). In fact, the correlations were statistically different. Whether directly related to the area of statistics or not, men’s attitude differs from that of women. Therefore, there is room to ask whether the performance of women does determine their level of confidence, or whether their level of confidence determines their performance. The nature of a possible intervention would be greatly influenced by the causal direction of these relationships. In the area of mathematics, the phenomenon of stereotype threat, where pre-existing negative stereotypes about one’s group can increase anxiety and, by extension, can decrease confidence in one’s abilities, would support the idea that confidence causally affects performance. This could be seen as being consistent with the fact that further education does not succeed in closing the gender gap.

In a related fashion, the second goal of this study was to examine the role of individual differences, first testing the appropriateness, for the area of statistical reasoning, of the tri-partite theory of reasoning proposed by Stanovich (2009). Stanovich’s argument relies on the idea that thinking dispositions motivate the use of cognitive ability to solve reasoning problems. Using a structural equation model to test the relation between thinking dispositions, cognitive ability, and statistical reasoning, the fit of the proposed model to the data was very good, and the pattern of relation between individual differences and statistical reasoning was equivalent across gender. Adding gender as a predictor in the model demonstrated how its influence on performance is complex, and multi-faceted. Indeed, gender is modeled as influencing statistical reasoning both directly – as demonstrated by the significant path between gender and statistical reasoning – and indirectly through its significant influence on thinking dispositions and on cognitive abilities. The subsequent studies will attempt to shed some light on possible factors at play in this equation, keeping in mind that the model indicates that any attempt to influence statistical reasoning directly or indirectly will likely address no more than one-third of the total effect of gender on statistical reasoning. Indeed, when examining the total effect composed by each of the three significant paths, one can see that 37% of the effect is explained by the influence that gender has on thinking dispositions; that 24% of the effect is explained by the influence that gender has on cognitive ability; and that 39% of the effect is explained by the direct influence of gender on statistical reasoning.

Just as it has been mentioned in mathematics, multiple factors should be considered when studying gender and performance, ranging from an individuals level of interest in the topic, to cognitive processes, to socialization (Byrnes, & Takahira, 1993). For instance, expectations and attitudes toward the self and toward others, including stereotypes, are acquired through socialization. When a stereotype suggests a negative characteristic about a group to which one identifies, that stereotype becomes threatening and can impede one’s ability to perform to its full potential. This stereotype threat hypothesis (Steele, 1997; Spencer et al., 1999) has been studied extensively in mathematics. The next set of studies will explore the validity of this hypothesis for the area of statistics.

STUDY 2

As shown in Study 1, individual differences in cognitive ability and in thinking dispositions are not sufficient to fully explain the performance gap in statistics. Gender influences statistical reasoning both in direct and indirect ways, and experience benefits both genders relatively equally. However, for females with comparable prior experience who are performing as well as males in class and obtaining the same grades, their performance on standardized tests such as the SAT-math (Byrnes & Takahira, 1993) and on statistical reasoning tests (Tempelaar et al., 2006) is nonetheless lower. Consistent with the model presented in Study 1, this finding makes it difficult to explain the gender gap in terms of pure ability. In fact, Byrnes & Takahira (1993) recommend that multiple factors be considered when studying gender and mathematics, including socialization. For instance, research has shown that prior beliefs and societal stereotypes – such as females not being good at math – are perpetuated by family and teachers alike and are difficult to eradicate from the classroom (Smith & Hung, 2008). They can also give rise to performance deficits (Shapiro & Williams, 2012), at least in part through the pervasive effect of stereotype threat. Stereotype threat occurs when members of a negatively stereotyped group anxiously expect that their performance will confirm the stereotype attached to their social group (Steele & Aronson, 1995; Spencer, Steele, & Quinn, 1999). This phenomenon has been documented in different areas, with different groups, such as intelligence testing in African-Americans (Steele & Aronson, 1995), athleticism in European Americans (Stone, Perry, & Darley, 1997), social sensitivity in men (Koening & Eagly, 2005), and mathematical abilities in women (Quinn & Spencer, 1999, 2001; Schmader, Johns, & Barquissau, 2004). In fact, gender is one of the most often cited sources of performance deficits related to stereotype threat.

Even though the phenomenon has not been established directly in statistics, the fact that statistics use mathematical tools (Moore, 1992), combined with the fact that females often show signs of anxiety when faced with either mathematics (Ashcraft, & Faust, 1994) or statistics (Onwuegbuzie & Wilson, 2003) classes, together make the field of statistics a likely candidate for stereotype threat. Moreover, Study 1 clearly shows that thinking dispositions influence statistical reasoning above and beyond differences in cognitive abilities. Reasoning theorists explain this phenomenon by emphasizing that, whereas cognitive ability is highly stable, thinking dispositions are malleable (Stanovich, 2001; Baron, 1985; Tishman, & Andrade, 1995), which makes thinking dispositions appropriate candidates for attempts at modification of behaviour. Theoretically, if we can influence females' willingness to engage with statistical material, we could indirectly improve their performance on the statistical reasoning task.

One way to encourage willingness to engage in a task is by manipulating the context of the task. For instance, in discussing Need for Cognition (a thinking disposition), the creators of the scale emphasize the stability of the trait while underlining how it is influenced by situational constraints (Cacioppo, Petty, & Kao, 1984). Therefore, assuming that this and other thinking dispositions can be influenced by the situation, and given that stereotype threat is inherent to the social situation, it is possible that by reducing the negative power of the stereotype threat, participants' tendency to think productively will be restored. If true, female participants – as they are the target of the negative stereotype - should have increased motivation to engage with the task and in turn perform better on the statistical reasoning task.

There is some evidence that explicit teaching about stereotype threat could help reduce its detrimental effects (Johns, Schmader, & Martens, 2005). Another potential strategy to help counter the effect of stereotype threat is through "value affirmation". Value affirmation is a social-psychological intervention that aims at increasing one's self-perception of worth (Yeager & Walton, 2011). By asking people to think about activities or values that are very important to them, value affirmation helps shift attention from the anxiety-inducing field or activity (Martens, Johns, Greenberg, & Schimel, 2006; Taylor & Walton, 2011). One main reason for the interest in this technique is the potential it holds as being easily employable by students prior to high-stake tests such as final exams, or at any point during learning.

A recent in-class initiative, testing the usefulness of value affirmation to counter stereotype threat in the male-dominated area of physics, showed some promise. Miyake, Kost-Smith, Finkelstein, Pollock, Cohen, and Ito (2010) randomly assigned students in an introductory physics class to a value affirmation group or to a control group. The manipulation was presented as a writing exercise that students had to complete mandatorily as part of the course. Both groups completed it, albeit on slightly different topics. In the control group, students were asked to write about a value of low importance to them. In contrast, the value affirmation group was asked to write about a value of high importance to them. Whereas males' performance did not differ based on the group to which they were assigned, females in the value affirmation group outperformed females in the control group. The positive impact of value affirmation has also been demonstrated for females in mathematics (Martens, Johns, Greenberg, & Schimel, 2006). Thus, value affirmation has been shown to be a worthwhile treatment to reduce gender gap in some academic settings. However, this idea has not yet been tested in the area of statistics.

In this study, we conducted a replication and extension of the Miyake et al (2010) study in the context of a statistics class. If stereotype threat is a viable explanation for the performance gap observed in statistics, then female students in a value affirmation condition should perform better than female students in a control condition on a statistical reasoning test at the end of the term, but not at the beginning of the term as the first testing round occurred prior to the value affirmation manipulation. The condition they are in should not affect the performance of male students. However, as found in the physics class (Miyake et al., 2010), it was possible that the level to which females endorse the stereotype may reveal itself as a moderator of the effect, with the value affirmation intervention proving beneficial only for those females who believe that females are less skilled than males in statistics. Thus, if stereotype threat does apply to the field of statistics, then performance should improve for females who self-affirm their values, yet perhaps only for those who believe that a negative stereotype overshadows their potential.

Method

Participants

Volunteers were recruited from an introduction to statistics course at the University of Waterloo (Psych 292 - "Basic Data Analysis") on January 4th 2011 (first day of class). Participants were invited to complete a 3-phase study. In phase 1, participants received chocolate and a chance to win one of five prizes of $20. In phase 2, participants received a chance to win one of three prizes of $30. In phase 3, participants received a chance to win one of two prizes of $50. Sixty-seven eligible students (17 males, 50 females) completed phase 1; thirty-one students (7 males, 24 females) completed phase 2; and twenty-three students (5 males, 18 females) completed all three phases.

Design

A 2 (gender) x 2 (condition) between-subjects design was used. Although experience was homogeneous across the group at each phase, its effect was tested within- subjects for those participants who completed all three phases. Males and females were randomly assigned either to the value affirmation or to the control condition. In each condition, as in Martens et al. (2006), participants first ordered a list of eleven values from the most important to them (#1) to the least important to them (#11) (see Appendix C). In the value affirmation group, participants were then asked to explain how and why the value they ranked #1 was important to them. In the control group, participants were asked to explain how and why the value they ranked #9 might be important to another UW student. Both the instructor and the teaching assistants were unaware of who participated in any of the phases, and were therefore blind to which condition any participant was in until the grades for the course had been officially submitted. All data was handled by a research assistant under the supervision of a faculty member, both otherwise uninvolved with the students of that course.

Materials

Five questionnaires were used in this study: (1) Statistical Reasoning Assessment (SRA: Garfield, 2003 – Appendix A); (2) the self-affirmation questionnaire (based on Martens et al, 2006 – Appendix C); (3) a stereotype endorsement scale adapted for mathematics and for statistics (based on Miyake et al., 2010 – Appendix B); (4) the Numeracy Scale (NUM: Lipkus et al., 2001; Coefficient alpha = ranging from 0.70 to 0.75 – Appendix B); and (5) the Preference for Numerical Information Scale (PNI: Viswanathan, 1993; Coefficient alpha = ranging from 0.94 – Appendix B). In addition, final grades from the pre-requisite to the introduction to statistics course (Psych 291 - “Basic Research Methods”) and final grades from the course (Psych 292 - “Basic Data Analysis”) were obtained for students who provided consent.

Procedure

Phase 1. Participants were recruited in class on the first day of the term. As per the ethics board guidelines, participation was voluntary. This had been established to prevent the students from feeling pressured to participate as the main researcher was also teaching this class. In phase

1, each participant received a chocolate and could enter a draw to win one of five prizes of $20. Participants were randomly assigned to one of the two conditions (i.e., affirmation vs. control). Prior to signing the consent form, they were informed that the study had three phases. They were asked for their consent for the current phase of the study, and for the permission to contact them by email to complete the next two phases online. They were also asked to grant us permission to access their final grades in the pre-requisite course in research methods (Psych 291) and in the current introductory statistics course (Psych 292).

During this phase, participants completed the SRA, to which the two stereotype belief questions had been added, followed by the value affirmation task. All participants completed the ranking of 11 values (see list in Appendix C) in the first place, followed by a short essay to complete afterward. Then again, the nature of the short essay varied, with one version focusing on why the value they ranked as #9 could be important to others, and the other version focusing on why the value they ranked as #1 is important to them. Completion of the tasks occurred in the classroom and took between 30 and 40 minutes.

Phase 2. Two weeks after Phase 1, all participants who had given their consent received an email and a link to take them to the Phase 2 of the study. In case someone's email address had been entered wrongly, a second appeal was made through the online posting board for the class, given that the sample size from the first phase was smaller than expected.

Phase 2 was included to emulate the general design used by Miyake et al. (2010), where the value affirmation task was repeated after a few weeks. Participants were informed that the task would occur online and take no longer than 10-15 minutes to complete. At the end of the task, participants could choose to enter a draw for one of three prizes of $30.

Phase 3. At the beginning of April 2011, thus three months after Phase 1, all participants who had previously given their consent received an email and a link to take them to the Phase 3 of the study. This phase required 30-45 minutes to complete and took place online. Participants were asked to complete three different questionnaires: (1) SRA (as their post-test), (2) NUM, and (3) PNI. At the end of the tasks, participants could choose to enter a draw for one of two $50 prizes. After all participants had completed the study, feedback was emailed to them.

Results

In this study, we were interested in examining the impact of value affirmation on the development of statistical reasoning. Data from phase 1 was analyzed as the baseline for the sample of 56 participants (out of 67) who gave access to their grade from the pre-requisite course, and data from phase 3 was analyzed as the post-test for the continuing 23 participants. It was expected that a gender effect would be present at baseline. However, if value affirmation is an effective intervention in statistics – implying the existence of a stereotype threat in that domain – females in the value affirmation condition should surpass those in the control condition at Phase 3, also reducing the gap between them and males in either condition. Indeed, it was not expected that condition would have an impact on the performance of males. Thus, the interaction between condition and gender should be significant. Descriptive statistics for correct reasoning, misconception and confidence can be found in Table 3 for Phase 1 and in Table 4 for Phase 3.

Table 3

Descriptive Statistics – Study 2 – Phase 1 – Performance & Confidence

	Male			**Female**			**Total**		
	n	*Mean*	*SD*	n	*Mean*	*SD*	n	*Mean*	*SD*
Correct reasoning	**16**	.69	.16	**40**	.60	.16	**56**	.63	.14
Misconception Scale		.21	.10		.30	.12		.27	.12
Confidence		4.92	.59		4.50	.90		4.62	.84

Table 4

Descriptive Statistics – Study 2 – Phase 3 – Performance & Confidence

		Male			**Female**			**Total**		
	Condition	n	*Mean*	*SD*	n	*Mean*	*SD*	n	*Mean*	*SD*
Correct reasoning Scale	Control	**3**	.81	.08	**10**	.63	.10	**13**	.67	.12
	Affirmation	**2**	.80	.04	**8**	.61	.17	**10**	.65	.17
	Total	**5**	**.81**	**.06**	**18**	**.62**	**.13**	**23**	**.66**	**.14**
Misconception Scale	Control		.14	.09		.24	.103		0.21	.10
	Affirmation		.12	.07		.33	.152		0.29	.16
	Total		**.14**	**.08**		**.28**	**0.13**		**0.24**	**.13**
Confidence	Control		5.13	.33		4.77	.79		4.85	.71
	Affirmation		5.35	.42		5.23	.37		5.26	.36
	Total		**5.22**	**.34**		**4.98**	**.66**		**5.03**	**.61**

Phase 1 – Pre-requisite, performance & confidence

Pre-requisite research methods (Psych 291) grade. To get a sense of the composition of the sample, grades from the pre-requisite course to the statistics course were obtained. An independent samples t-test demonstrated equal performance across gender in the basic research course, despite females (*M* = 85.80, *SD* = 8.68, with grades ranging from 60 to 100) scoring even slightly higher than males (*M* = 82.63, *SD* = 10.84, with grades ranging from 55 to 95), although not significantly so, *t*(54) =1.15, *SE* = 3.18, *p* = .26.

Performance. In contrast, females underperformed on the statistical reasoning assessment, *t*(54) = 2.15, *SE* = .041, *p* = .036, scoring 9% less than males on the correct reasoning scale (M_f = .60, *SD* = .13; M_m = .69, *SD* = .16), and adhering to 9% more misconceptions than males, *t*(54) = 2.76, *SE* = .03, *p* = .008, (M_f= .30, *SD* = .12; M_m = .21, *SD* = .10). Thus, despite an equal performance on a pre-requisite for a statistics course, the gender gap easily appears when the same group is submitted to a statistical test.

In this sample, 64% of participants disagreed with the gender stereotype in statistics. In other words, only 25% of males and 40% of females agreed with the potential stereotype in statistics that men generally do better in statistics than women. Using the level of endorsement of the stereotypes as a covariate (*p* = .94) in the general linear model to analyze correct reasoning left the gender effect intact (*p* = .041). Finally, controlling for the grade obtained in the prerequisite course in research methods (Psych 291) – a significant covariate (*p* = .001) – did not influence the obtained gender effect (*p* = .006). The same patterns of results were found when analyzing the misconceptions scores with the same covariates. Summary tables of results for the ANCOVA analyses are presented in Appendix D.

Confidence. Reflecting the results on the performance measure, females were less confident than males (M_f= 4.50, SD = .90; M_m = 4.92, SD = .59), even though the difference did not quite reach significance, $t(54) = 1.68$, $SE = .25$, $p = .10$. Unlike Study 1, males' and females' confidence was calibrated with their performance [$r(14) = .56$, $p = .024$ and $r(38) = .33$, $p = .038$, respectively]; however, their degree of calibration did not differ from each other ($z = .9$, $p = .37$). A scatterplot of the relation between performance and confidence is presented in Figure 5, and the complete set of correlations is available in Table 5.

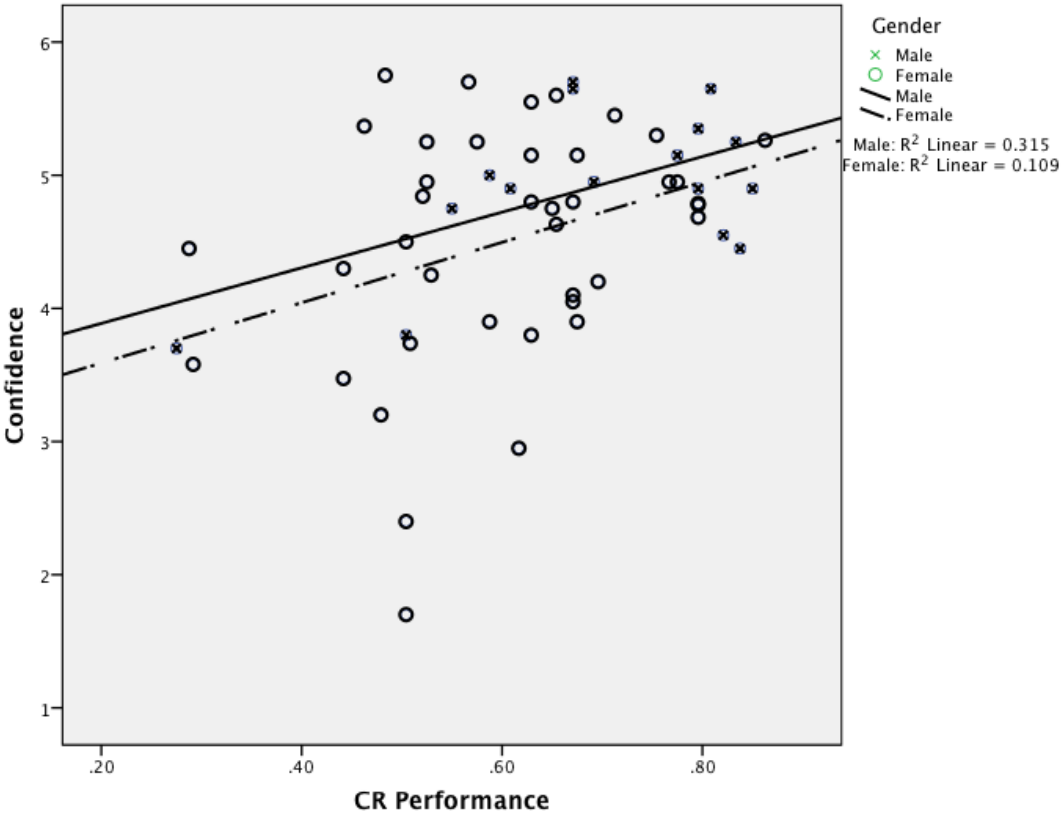


Figure 5. Scatterplot of the association between performance and confidence, by gender. Study 2, Phase 1.

Table 5

Correlation Matrix – Study 2 – Phase 1

Subscale	**Mean (SD)**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. CR	.63 (.14)	--	$-.62^{**}$	$.41^{**}$	-.05	-.15	$.38^{**}$	$.30^{*}$
2. MISC	.27 (.12)		--	$-.27^{*}$	.05	.11	-.17	-.11
3. Confidence	4.62 (.84)			--	-.09	-.03	.20	.18
4. SE_Stats	2.80 (1.43)				--	$.95^{**}$	-.04	-.19
5. SE_Math	3.05 (1.65)					--	-.05	-.16
6. Pre-req.grade	84.89 (9.36)						--	$.60^{**}$
7. Course grade	80.91 (10.86)							--

Note. * $p < .05$. ** $p < .01$.

Phase 3 – Performance & confidence

For this phase, the value affirmation condition can be used as an independent variable of interest. Performance scores and confidence ratings were analyzed using a 2 (gender) x 2 (condition) analysis of variance. When appropriate, a covariate was added to this general linear model. The expectation was that females in the value affirmation condition would perform better than females in the control condition, and that males' performance and confidence would not be affected by the condition, thus leading to a gender by condition interaction. Descriptive statistics for all three dependent variables can be found in Table 4.

Performance. As with Phase 1, a gender effect was present on both reasoning scales (**CR**: $F(1, 19) = 8.46$, $MSE = .016$, $p = .008$, $\eta^2_p = .308$; **MISC**: $F(1, 19) = 5.62$, $MSE = .029$, $p = .029$, $\eta^2_p = .228$), with females underperforming on the correct reasoning scale by 19% ($M_f = .62$, $SD = .13$; $M_m = .81$, $SD = .06$), and adhering to significantly more misconceptions than males by 14% (M_f= .28, *SD* = .13, M_m= .14, SD=.08). Unfortunately, the value affirmation condition did not help improve the performance of women on the SRA, neither as a main effect of condition, (**CR**: *F*(1, 19) = .04, *MSE* = .016, *p* = .853, η^2_p = .002; **MISC**: *F*(1, 19) = .31, *MSE* = .015, *p* = .587, η^2_p = .016) nor a gender by condition interaction, (**CR**: *F*(1, 19) = .01, *MSE* = .016, *p* = .912, η^2_p = .001; **MISC**: *F*(1, 19) = .82, *MSE* = .015, *p* = .377, η^2_p = .041), was present. This pattern of effects held true even when controlling for the level of endorsement (**CR**: *p* = .24; **MISC**: *p* = .75) of the gender stereotype in statistics (see Appendix D for summary tables of these and other ANCOVA analyses). Condition was again not significant when limiting the analysis to females (**CR**: *t*(16)= .31, *SE* = .063, *p* = .76; **MISC**: *t*(16)= 1.51, *SE* = .060, *p* = .15). Also, when splitting the group by whether participants agreed or disagreed with the stereotype, as done in Miyake et al. (2010), although the effect of condition was significant for the correct reasoning, *F*(1, 5) = 11.746, *MSE* = .002, *p* = .019, η^2_p = .701, the overly small sample size of eight, which includes only one female in the affirmation condition, prevented drawing any meaningful conclusion.

Confidence. Despite females having a slightly lower level of confidence than males (M_f = 4.98, *SD* = .66; M_m = 5.22, *SD* = .34), and despite those in the affirmation condition having a slightly higher level of confidence ($M_{affirmation}$ = 5.26, *SD* = .36; $M_{control}$ = 4.85, *SD* = .71), neither gender, *F*(1, 19) = .60, *MSE* = .366, *p* = .448, η^2_p = .031, condition, *F*(1, 19) = 1.19, *MSE* = .366, *p* = .289, η^2_p = .059, nor their interaction, *F*(1, 19) = .16, *MSE* = .366, *p* = .699, η^2_p = .008, had a statistically significant effect on confidence. Also, males' and females' confidence was not significantly calibrated with their performance (*r*(3) = .57, *p* = .319 and (*r*(16) = .10, *p* = .695. respectively); possibly due to the overly small sample, their degree of calibration did not differ from each other (*z* = .72, *p* = .47). A scatterplot of the relation between performance and confidence is presented in Figure 6, and the complete set of correlations is available in Table 6.

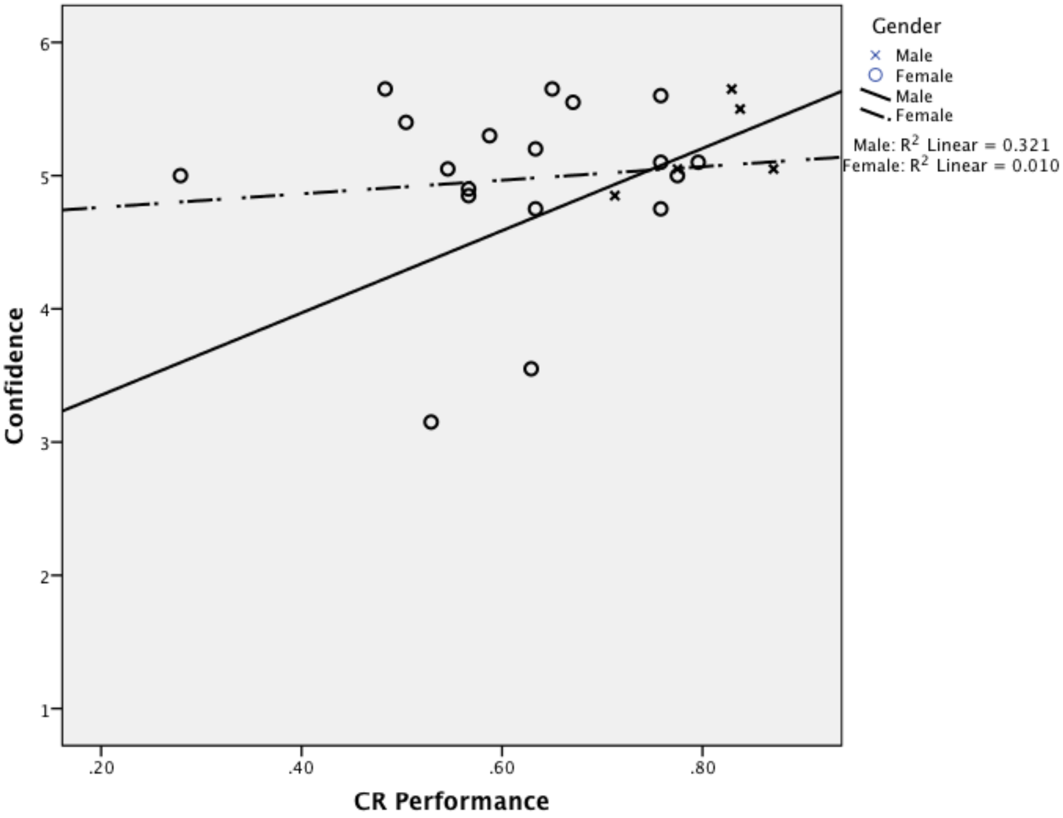


Figure 6. Scatterplot of relation between performance and confidence, by gender. Study 2, Phase 3.

Table 6

Correlation Matrix – Study 2 – Phase 3

Subscale	**Mean (SD)**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1. CR	.65 (.14)	--	-.66**	.20	.21	.38	-.09	-.34	.36	.07	-.47*
2. MISC	.25 (.11)		--	-.21	-.20	-.24	-.03	.12	-.14	-.03	.33
3. Conf	4.95 (.64)			--	.34	.30	-.29	-.19	-.18	-.06	-.08
4. PNI	83.35 (12.42)				--	.39	-.42*	-.45*	.32	.16	.02
5. NUM	8.57 (1.65)					--	-.01	-.03	.45*	.47*	-.01
6. STstats	2.83 (1.44)						--	.91**	-.13	.19	-.09
7. STmath	3.22 (1.70)							--	-.25	.19	.05
8. 291	86.90 (8.11)								--	.60**	-.14
9. 292	84.40 (9.32)									--	.15
10. Word Count	80.09 (31.50)										--

Note. * *p* < .05. ** *p* < .01.

Discussion

In this study, the gender gap observed in Study 1 was replicated. As the grades of students were not different on the pre-requisite course, this gender gap on the SRA could be seen as somewhat surprising. Then again, Tempelaar et al. (2006) also found the gender gap despite the equivalent background experience of males and females. Importantly, no effect of value affirmation was found, and taking participants' belief in the gender stereotype into account did not moderate its effect. Although not statistically significant, males demonstrated a trend toward higher confidence than women in Phase 3. However, value affirmation did not yield higher confidence. Also, unlike Study 1, the calibration of males and females did not differ. Given the small sample sizes, this lack of significant differences between correlation coefficients is not surprising.

The finding of lack of effect of the technique of value affirmation is disappointing, but there are a few reasons why it may have been ineffective in this case. First, it is possible that the sample was atypical given that participants were self-selected. Unlike Miyake et al. (2010), we were not given the permission to enlist the entire class in the research project. Despite the fact that grades from the pre-requisite course ranged from 55 to 100 for the sample in Phase 1 and ranged from 71 to 100 for the sample in Phase 3, the group average was around 85% in each phase, which supports the idea that the sample contained higher-performing students. In fact, their average is higher than the overall group average of 79% in the pre-requisite course for their cohort. Thus, it is possible that this high-performing group could have been less susceptible to the self-affirmation manipulation. Another sign of their distinction from the norm comes from comparing SRA scores in Study 2 to those in Study 1. At Phase 1, females scored 11% higher than females with no experience in statistics in Study 1, and males scored 12% higher than males with no experience in statistics in Study 1. At Phase 3, females scored 7% higher than females with one course in statistics, and males scored 14% higher than males with one course in statistics.

A second possibility to explain the lack of effect of value affirmation is that statistics is not processed the same way as mathematics. Indeed, statistics educators strongly argue that statistics are fundamentally different from mathematics (Moore, 1992; Iversen, 1992). Even though statistics uses the mathematical language, just as economics and physics do, it deals with different issues from mathematics, especially that of uncertainty (Iversen, 1992). In fact, statistics is the science of data and it deals specifically with numbers *within contexts* (Moore, 1992). Therefore, if statistics is not technically a branch of mathematics, this could suggest that stereotype threat may not be at play in statistics. However, this was not evident from participants' responses in the current study as no difference between the level of endorsement of the stereotype in mathematics and the level of endorsement of the stereotype in statistics was observed. The levels of endorsement in the current study were also similar to those reported by Martens et al. (2006).

Third, the class context may have played a role. Stereotype threat is often associated with being a minority (Steele, 1997; Shapiro & Williams, 2012). However, in the context of this statistics class for psychology majors, women are in the majority (i.e., 69% were women). Nevertheless, this may be offset by the fact that such negative stereotypes would be pernicious, especially on a campus populated greatly by the male-dominated mathematics, engineering, and computer science majors.

Finally, the most likely reason for this lack of a positive effect of value affirmation is the very limited sample size and the related lack of power. This project began ambitiously with the goal of replicating Miyake et al. (2010) in a statistics class. Ethical restrictions, especially the impossibility of making participation mandatory, prevented this project from achieving its entire potential. Although repeating this study in class was not an option, a replication was conducted in the lab. This allowed the recruitment of a larger sample and, as in Study 1, the inclusion of experience as a factor.

STUDY 3

While it was not possible to repeat the in-class experiment, Study 3 maintained the goal of testing the usefulness of value affirmation in statistics. Using a larger sample also provided the opportunity to reprise experience as a between-subject factor. In addition to allowing a replication of Study 1 (due to the inclusion of Gender *and* Experience as independent variables), it was possible to verify if experience moderates the usefulness of value affirmation for females.

If a stereotype threat does exist in statistics, not only is it expected that the gender gap will again be found, with males outperforming females, but taking action to reduce the impact of the stereotype threat should help improve performance of women in statistics. Specifically, females who affirm their values should perform better than females who do not affirm their values. However, given that value affirmation for those who are not under the threat of the negative stereotype (i.e., males in statistics) should have limited or no influence on their performance, we expected to find an interaction between gender and condition, with only females benefiting from the value affirmation manipulation.

In addition, as with the previous studies, confidence was measured to complement the performance-related results. In general, if participants are well calibrated, their confidence ratings should correlate positively with their performance. Given the results of Study 1, males were expected to report high confidence regardless of their performance, whereas women's confidence level should track their performance more closely. Although this pattern was not found in Study 2, the limitation in power discussed earlier is nonetheless a likely reason for this failure to replicate. Then again, if value affirmation does indeed reduce the fear of confirming the negative stereotype, females in the affirmation condition may also display higher confidence than women in the control condition.

Method

Participants

One hundred and fifty-seven undergraduate students (69 males, 88 females; M_{age} = 20.31, *SD* = 3.29) from the University of Waterloo were recruited through SONA, an experiment management system, and participated for course credit in any eligible psychology course. Participants had varying levels of experience in statistics, as measured by the number of statistics courses they had previously taken (0: N = 79; 1: N = 49; 2 or more: N = 29)[3].

Design & Materials

For this study, a 2 (gender) x 2 (condition) x 3 (experience) between-subjects design was used. As in Study 1, condition was randomly assigned and determined by the version of the value affirmation exercise that participants completed (see Appendix C). As with Studies 1 and 2, the SRA was used, and the dependent variables included correct reasoning scores (CR) and misconception scores (MISC), as well as confidence ratings.

The level of endorsement of the stereotype in statistics and in mathematics was measured. Unlike Study 2, the two questions were added at the end of the PNI scale. Although this meant that the rating would now be given on a 7-point rather than on a 6-point scale, it seemed more

[3] Despite conducting testing for this study for two terms, it proved difficult to recruit participants having taken at least two courses in statistics. Given that 29 is quite small in comparison to 79 in the 'stats = 0' condition, analyses were repeated after combining all participants with any level of experience within one group of 78 participants. This 2 (gender) x 2 (condition) x 2 (experience) ANOVA revealed the same two effects for correct reasoning performance as when three levels of experience were used. Specifically, females (*M* = .54, *SD* = .16) underperformed as expected in comparison to males (*M* = .62, *SD* = .16), *F*(1, 149) = 9.37, *MSE* = .025, *p* = .003, η^2_p = .059, and gender interacted with experience, *F*(1, 149) = 6.24, *MSE* = .025, *p* = .014, η^2_p = .040.

When reasoning turns to avoiding misconceptions, the results of the omnibus ANOVA revealed the same main effect of experience, *F*(1, 149) = 6.44, *MSE* = .014, *p* = .012, η^2_p = .041. Only one small difference emerged in this version of the analysis, with gender interacting with experience, *F*(1, 149) = 4.25, *MSE* = .014, *p* = .041, η^2_p = .028, as females benefited more greatly from experience (9% average decrease in misconceptions) than males (1% average decrease in misconceptions).

Finally, analysis of the confidence ratings revealed the same experience by condition interaction, *F*(1, 149) = 5.02, *MSE* = .494, *p* = .027, η^2_p = .033. Overall, using 3 levels versus 2 levels of experience in the analysis made very little difference.

natural and less disruptive to ask for participants' beliefs at the end of a self-report questionnaire rather than at the end of the assessment tool. The level of endorsement of the stereotype in statistics was subsequently used both as a continuous predictor in the general linear model, and as a basis to split the group to better examine whether believing in the stereotype moderated the effect of the value affirmation exercise.

Along with the other covariates presented in Appendix D (e.g., CRT, PNI), the number of words participants used in the writing exercise was counted and used as an additional covariate. This was done to control for the possibility that one's level of involvement in the task could influence the effectiveness of the value affirmation exercise.

Procedure

Participants came to the lab for this study and were tested in groups of one to three. After reading the informed letter and signing a consent form, participants were asked to complete four tasks: (1) the value affirmation task, (2) the SRA with its associated confidence ratings, (3) the CRT, and (4) the PNI, which ended with the two questions about their level of belief in a gender stereotype in math and in stats. At the end of the session, participants had the opportunity to ask questions about the study and received a feedback form.

Results

Performance and confidence scores were analysed using a 2 (gender) x 2 (condition) x 3 (experience) between-subjects ANOVA. It was expected that, as in Study 1, males would perform better than females and that experience would be equally beneficial to both genders. If the value affirmation exercise is effective at combating stereotype threat in statistics, women in the value affirmation condition should score higher than those in the control condition. However, this may only hold for women who believe in a stereotype in statistics. Based on the findings from Study 1, confidence was expected to be high and homogeneous for males irrespective of performance, while it was expected that confidence would be related to performance for women. In addition, if value affirmation does indeed reduce stereotype threat, females in the affirmation condition may also display higher confidence than women in the control condition. Prior to running the analyses, data exploration identified one extreme outlier that was removed from the analyses involving confidence ratings. The sample size for the subsequent analyses involving confidence scores was thus 156. The descriptive statistics for this sample are presented in Table 7.

Performance. A first look at the results of the omnibus ANOVA for correct reasoning performance revealed a single main effect of gender, $F(1, 145) = 4.60$, $MSE = .025$, $p = .034$, $\eta^2_p = .031$, with females underperforming as expected in comparison to males ($M_f = .54$, $SD = .161$; $M_m = .62$, $SD = .159$). The only other significant effect in this analysis was the interaction of gender by experience, $F(2, 145) = 3.15$, $MSE = .025$, $p = .046$, $\eta^2_p = .042$. To better understand this effect, the data file was split by gender. The follow-up one-way ANOVA revealed no significant effect of experience for males, $F(2, 82) = 1.68$, $MSE = .025$, $p = .19$, whereas the same analysis revealed a marginally significant effect of experience for females, $F(2, 82) = 2.95$, $MSE = .025$, $p = .058$, where only those with the most experience scored better than the group with no experience ($p = .057$ using a conservative *Tukey HSD* test, or $p = .022$ using *Fischer's LSD,* which is the most powerful while keeping the alpha level at .05 when only three groups are compared; Howell, 2007).

Table 7

Descriptive Statistics – Study 3 – Performance & Confidence

			Male			**Female**			**Total**	
	# Stats courses taken	n	*Mean*	*SD*	n	*Mean*	*SD*	n	*Mean*	*SD*
Correct reasoning Scale	0	**35**	.65	.14	**44**	.51	.15	**79**	.57	.16
	1	**22**	.57	.15	**27**	.56	.17	**49**	.57	.16
	2^+	**12**	.65	.21	**17**	.61	.14	**29**	.63	.17
	Total	**69**	**.62**	**.16**	**88**	**.54**	**.16**	**157**	**.58**	**.16**
Misconception Scale	0		.27	.13		.35	.114		.31	.13
	1		.29	.12		.28	.111		.28	.12
	2^+		.22	.13		.24	.097		.23	.11
	Total		**.27**	**.13**		**.30**	**.118**		**.29**	**.12**
Confidence	0		4.88	.59	**(43)**	4.78	.64		4.82	.61
	1		4.69	.74		4.79	.57		4.75	.65
	2^+		4.95	.76		4.96	.64		4.96	.68
	Total		**4.83**	**.67**	**(87)**	**4.82**	**.62**	**(156)**	**4.82**	**.64**

The non-significant effect of condition could be a sign that, as the results from Miyake et al. (2010) suggest, the value affirmation procedure is only effective for those who agree with the stereotype. To account for this possibility, the omnibus analysis was repeated, this time controlling for participants' level of endorsement of the stereotype in statistics. Although the overall pattern of results of the analysis did not change, stereotype endorsement appeared as a significant covariate (*p* = .025) (see detailed results of this and other ANCOVA analyses in Appendix D). Subsequently, participants' levels of endorsement of the stereotype were used to create two groups: one group in disagreement with the stereotype, and one group in agreement with the stereotype (with the neutral answer being integrated with the "agree" group in an attempt to equalize sample sizes)[4]. As in Study 2, a wide majority of participants (65%) disagreed with the stereotype in statistics. However, this time, more males (43%) than females (28%) agreed with the stats stereotype. This is virtually identical to the levels of endorsement of the stereotype in mathematics by the same sample (64% disagree; 43% of males and 31% of females agree). For the next analysis, to avoid creating very small cells (e.g., $n = 2$), the experience variable was dropped, and the sample was divided by endorsement type (agree vs. disagree). Conducting the same 2 (gender) x 2 (condition) ANOVA separately for each group, the results did not replicate Miyake's findings. Indeed, for both groups, although critically for the "agree" group, the affirmation condition did not lead to an improvement in performance. However, the habitual gender gap was present ["disagree": $F(1, 98) = 6.33$, $MSE = .027$, $p = .014$, $\eta^2_p = .061$; "agree": $F(1, 51) = 4.76$, $MSE = .023$, $p = .034$, $\eta^2_p = .085$] and was not qualified, in either group, by an interaction of condition with gender.

When reasoning turns to avoiding misconceptions, the results of the omnibus ANOVA revealed a single main effect of experience, $F(2, 145) = 4.98$, $MSE = .014$, $p = .008$, $\eta^2_p = .064$. Specifically, one course is not enough. As found in Study 1, it takes at least two courses in statistics to witness a significant decrease in the number of misconceptions that one uses (*Tukey HSD*, $p = .004$; $M_0 = .31$, $SD = .13$; $M_1 = .28$, $SD = .12$; $M_{2+} = .23$, $SD = .11$). No other effect was present. In addition, the level of belief in the stereotype was not a significant covariate ($p = .24$),

[4] If those with a neutral answer are removed, 130 participants are included in the analysis instead of 157. Of those 130, only 28 (22%) of those agreed with the stereotype. For all three dependent variables (CR, MISC, confidence), the effects of gender, condition, and their interactions were not significant. Except for the effect of gender in regard to CR [$F(1, 24) = 2.50$, $MSE = .030$, $p = .127$, $\eta^2_p = .094$] and MISC [$F(1, 24) = 1.36$, $MSE = .030$, $p = .255$, $\eta^2_p = .054$], all other F values were less than 1. Of course, this lack of effects is compounded by the fact that so few participants were now included in the analyses, with all four cells including less than 10 participants.

and its inclusion did not change the pattern of results. The 2 (gender) x 2 (condition) ANOVA on each endorsement group corroborated this, as all effects were non-significant (all six *F*s < 1, see Appendix D).

Confidence. Confidence ratings were analyzed using a 2 (gender) x 2 (condition) x 3 (experience) ANOVA. A single interaction effect of experience by condition was present, *F(2, 144)* = 3.10, *MSE* = .402, *p* = .048, η^2_p = .041. To better understand this effect, the sample was split by condition and the effect of experience was further analyzed. Whereas confidence did not change with experience in the control group, for the group who affirmed their values, those with the most experience had higher levels of confidence than those with no experience in statistics (*p* = .092 when the conservative *Tukey HSD* is used, but *p* = .037 when *Fischer's LSD* is used, as it is an appropriate choice when no more than three groups are being compared).

Endorsement level, used either as a covariate (*p* = .65) or to split the group, did not change the pattern of results. Mirroring performance, agreeing with the stereotype and affirming one's values did not have any beneficial effect on confidence. The full set of analyses (ANCOVA and split groups included) is available in Appendix D.

Finally, the calibration of males (*r* = .20, *p* = .105) was compared to that of females (*r* = .33, *p* = .002), but the difference found in Study 1 was not replicated (*z* = .85, *p* = .40). Here, males and females were comparably calibrated. A scatterplot of those correlations is presented in Figure 7. All other correlations are available in Table 8.

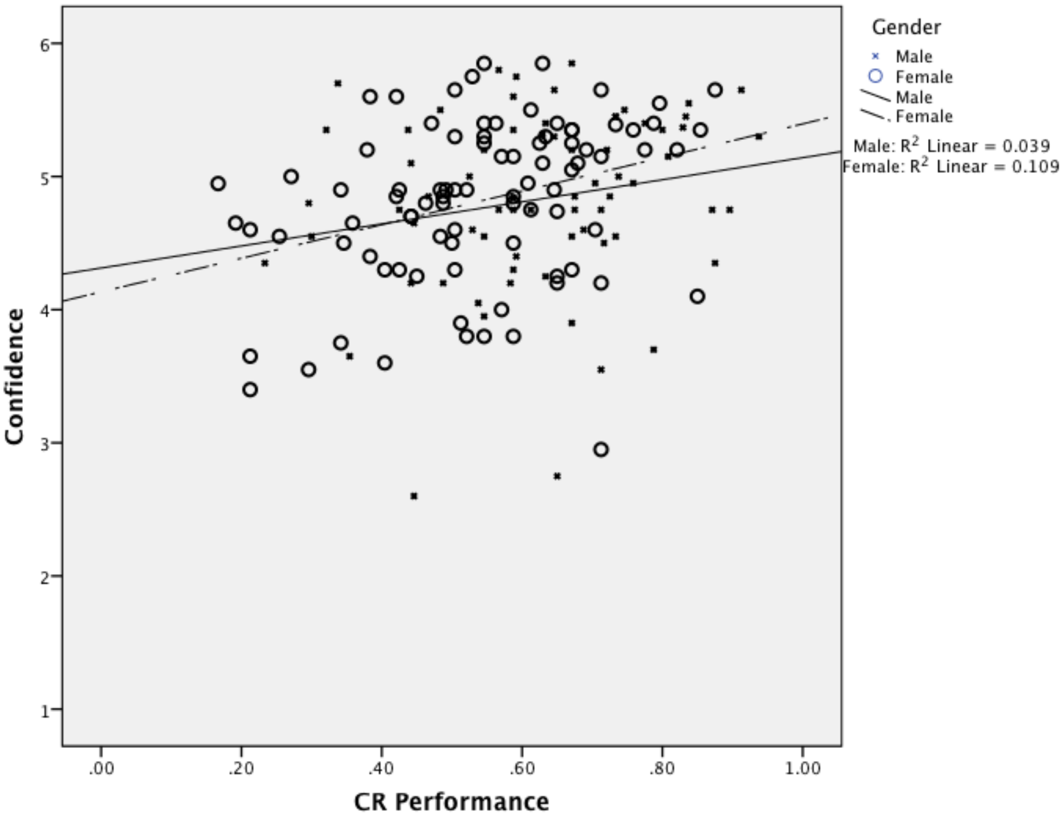


Figure 7. Scatterplot of association between performance and confidence, by gender. Study 3.

Table 8

Correlation Matrix – Study 3

Subscale	**Mean (SD)**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1. CR	.58 (.16)	--	-.65**	.26**	.33**	.45**	-.14	-.14	-.01
2. MISC	.29 (.12)		--	-.27**	-.32**	-.38**	.07	.10	.06
3. Confidence	4.82 (.64)			--	.14	.06	-.05	-.02	-.04
4. PNI	92.75 (15.24)				--	.47**	-.02	.01	-.09
5. CRT	1.45 (1.15)					--	.08	.10	.04
6. STstats	2.70 (1.67)						--	.88**	.16*
7. STmath	2.73 (1.77)							--	.15
8. Word Count	60.59 (32.63)								--

Note. All correlations involving confidence are based on N=156 instead of N=157.
* $p < .05$. ** $p < .01$.

Discussion

In this study, the gender gap was once again evident on the correct reasoning performance, whereas experience was found to be useful in reducing – but not fully eliminating – misconceptions. Using a larger sample, it was hoped that value affirmation would reveal itself as a useful tool to improve performance on a statistical reasoning test. Even for the group who believed in the stereotype, this was not the case. As in Study 2, affirming one's prized values had no impact on reasoning or confidence.

There are a few ways in which this lack of effect could be explained. First, as mentioned in Study 2, it is possible that statistical reasoning is different from mathematical reasoning, and that this domain is not susceptible to the negative influence of a stereotype. The finding that those who agree with the stereotype do not benefit from value affirmation supports this idea. In addition, everyone does not consider statistics a subfield of mathematics. Moore (1992) is the great defender of the idea that statistics, despite using tools from the area of mathematics, is a separate field. Statistics are not simply calculations but rather the science of data – of making sense of numbers presented in a specific context. Onwuegbuzie and Wilson (2003) also argue that the assumption that statistics and mathematics are the same is outdated. Since statistical software became mainstream in the early 1990s, statistics no longer requires long mathematical equations to be solved by hand. The focus of statistics is now on big ideas rather than on heavy calculations. As such, these authors suggest that the literature needs to adapt and stop equating statistics with mathematics.

Second, some will question the choice of not making the stereotype more salient, by asking participants to report their gender at the beginning of the study, for instance. Sackett et al. (2004, 2008) have questioned the external validity of the findings of a stereotype threat in the laboratory. As such, there is a need for studies that examine the stereotype threat hypothesis in a more natural setting. Given that, in real-life, the negative stereotype surrounding the ability of females in mathematics is pervasive and communicated through a variety of sources (Walton & Spencer, 2009), it seemed appropriate to capitalize on this pervasiveness alone to improve the external validity of the findings. In addition, literature on the stereotype threat in mathematics also notes that the stereotype does not need to be made explicit to take its toll (Steele, 1997; Spencer et al., 1999). Furthermore, the importance of statistical reasoning is obvious in everyday life, both in personal and in work spheres. In those contexts, women are not repeatedly asked to report their gender prior to making decisions, nor should they be. Also, despite not emphasizing the inherent threat of the situation in the design of the study, the gender gap in performance is apparent. Thus, something is at play in creating the gender gap. It is simply not convincing yet that stereotype threat is the factor of interest.

Prior to accepting that value affirmation is not useful in reducing the gender gap in statistics, there is room to question the testing instrument. The SRA is a fairly short test that represents a limited number of topics (Garfield & Chance, 2000), and which has been criticized for focusing too much on the topic of probabilities (Garfield, 2003). After the creation of and some work with the SRA, Garfield and her colleagues (delMas, Garfield, Ooms, & Chance, 2007) proposed a new test. The Comprehensive Assessment of Outcomes in Statistics (CAOS) comprises many of the questions of the SRA, also uses a multiple-choices format, and shares the same goal of evaluating knowledge related to a first course in statistics. However, it covers a wider range of topics and focuses less on probability. In addition, the internal consistency of the test is very good (DeVellis, 1991) with a Cronbach's alpha coefficient of 0.82, which is higher than the reliability reported by Garfield (2003) for the SRA (Cronbach's alpha coefficient equal to .70 for the correct reasoning scale and to .75 for the misconception scale). For those reasons, the next study used the same design; however, I used the CAOS as the primary measure of statistical competence, rather than the SRA.

STUDY 4

Given the absence of a positive effect of the value affirmation manipulation in the previous study, it was decided to replace the statistical reasoning task to verify if this lack of finding is simply contextual and linked to the test material being used. Therefore, in this fourth study, the aim is to examine whether the test used previously can account for not finding the expected effect of the value affirmation manipulation in statistics. To do so, I administered the CAOS instead of the SRA. As mentioned earlier, the CAOS is a test that is comprised of 40 multiple-choice questions, many of which were part of the SRA. The goal of the CAOS is also to evaluate the knowledge students should possess after a first course in statistics (see Appendix E for a list of the learning objectives associated with each question). In addition, the reliability of the CAOS is higher than that on the SRA (delMas, Garfield, Ooms, & Chance, 2007). These similarities and improvements make the CAOS an appropriate replacement choice for an attempted replication and extension of the previous study.

Logistically, given that the CAOS includes 40 questions instead of 20, it was also decided to modify the method for the collection of confidence ratings to avoid increasing the total length of the session. Therefore, instead of asking participants to report their level of confidence for each question, participants were asked at the end of the test to report what percentage of the questions they believed to have gotten right. In addition to keeping the testing time reasonable (and comparable to prior experiments), this change removes the possibility that confidence ratings could be distorted over time from the act of repetitively having to report them after each problem. The percentage scale may also reveal itself to be a more sensitive tool to measure the impact of value affirmation.

If value affirmation is sensitive to the test material used, we expect that performance on this new test could improve for those in the value affirmation condition, although possibly only for those who agree with the stereotype. On the other hand, if stereotype threat is not a reality in the domain of statistics, then we expect that the value affirmation manipulation will have no effect despite using a new test. If this alternative format for probing confidence (i.e., only once after the test is done) is more sensitive than asking for a rating after each question, then condition may, for once, appear to increase confidence for those females in the affirmation group, especially if they believe in a stereotype in statistics.

Method

Participants

Two hundred and six undergraduate and graduate students (97 males, 109 females; M_{age} = 21.33, *SD* = 4.56) from the University of Waterloo participated for course credit or remuneration. Participants had varying levels of experience in statistics, as measured by the number of statistics courses they had previously taken (0: N = 73, 1: N = 68, 2^{+}: N = 65), and came from fields deemed generally non-quantitative to extremely quantitative [Generally non-quantitative = 13 (6%), Minimally quantitative = 25 (12%), Moderately quantitative = 82 (40%), Highly quantitative = 69 (34%), Extremely quantitative = 15 (7%)].

Design & Materials

As in study 3, a 2 (gender) x 2 (condition) x 3 (experience) between-subjects design was used. Given that the CAOS was used, the number of dependent variables was reduced to two: correct reasoning performance and confidence. Indeed, the CAOS does not include a misconception subscale, making it a simpler instrument to use and score. Participants' levels of endorsement in the gender stereotype in statistics and in mathematics were collected to use as covariates. Other covariates included the CRT (as a proxy for cognitive ability), PNI (as a proxy for thinking dispositions), and the number of words used in the writing exercise (as a proxy for involvement in the task). Except for the analysis of covariance using the level of endorsement of a statistics stereotype, all other analysis results are presented in Appendix D to avoid overcrowding the results section.

As with Studies 2 and 3, participants were randomly assigned to a condition, which was determined by the version of the value affirmation exercise that participants completed. As before, all participants first completed the ranking of 11 values (see list in Appendix C), followed by a short essay. However, the nature of the short essay varied, with one version focusing on why the value they ranked as #9 could be important to others, and the other version focusing on why the value they ranked as #1 is important to them.

Procedure

This study took place in a lab. After reading the informed consent letter and signing a consent form, participants were asked to complete (1) the value affirmation task, (2) the CAOS with its associated confidence rating, (3) the CRT, (4) the PNI along with the two stereotype endorsement questions, as well as (5) the demographic information based on Schield (2005). At the end of the session, participants had the opportunity to ask questions about the study and received a feedback form.

Results

As in Study 3, performance and confidence scores were analyzed using a 2 (gender) x 2 (condition) x 3 (experience) between-subjects ANOVA, with the difference that only one correct reasoning score captured performance. It was expected that, as in all three previous studies, males would perform better than females and that experience would be equally beneficial to both genders. If the value affirmation exercise is effective at combating stereotype threat in statistics, then women in the value affirmation condition should score higher than those in the control condition. However, this pattern may only be evident for those who agree with the gender stereotype in statistics.

For women, it is expected that confidence would be positively related to performance. In addition, if value affirmation does indeed reduce stereotype threat and improve performance, females in the affirmation condition may also display higher confidence than women in the control condition. For men, however, the pattern emerging from Studies 1 and 3 is not as clear. Thus, their confidence could be high and homogeneous irrespective of performance (as in Study 1), or it could follow performance (as in Studies 2 and 3). At the very least, men's confidence should be higher than women's confidence level. Descriptive statistics for performance and confidence are available in Table 9.

Performance

Performance on the CAOS test was first analyzed using 2 (gender) x 2 (condition) x 3 (experience) between-subjects ANOVA. The results replicated Studies 1, 2 and 3, with a significant gender effect where females (M = .48, SD = .108) score 12% lower than males (M = .56, SD = .143), F(1, 194) = 17.62, MSE = .015, p < .001, η^2_p = .083, and a significant effect of experience, F(2, 194) = 6.87, MSE = .015, p = .001, η^2_p = .066. Specifically, participants with any level of experience perform better than those with no experience in statistics (*Tukey HSD*, p < .015). Despite those with the most experience (M_{2+} = 0.56, SD = 0.14) scoring higher than those with only one course in statistics (M_1 = 0.53, SD = 0.12), the difference was not statistically significant (*Tukey HSD*, p > .05). Noticeably again, the effect of condition was not significant, and entering the level of belief in the stereotype as a covariate (p = .20) did not change the

Table 9

Descriptive Statistics – Study 4 – Performance & Confidence

		Male			**Female**			**Total**		
	# Stats courses taken	n	*Mean*	*SD*	n	*Mean*	*SD*	N	*Mean*	*SD*
Performance – CAOS	0	**30**	.54	.12	**43**	.42	.09	**73**	.47	.12
	1	**32**	.55	.13	**36**	.52	.11	**68**	.55	.13
	2^+	**35**	.60	.17	**30**	.52	.10	**65**	.56	.14
	Total	**97**	**.56**	**.14**	**109**	**.48**	**.11**	**206**	**.52**	**.13**
Confidence	0		61.28	19.42		56.60	17.08		58.53	18.10
	1		57.68	22.29		57.61	17.61		66.55	17.70
	2^+		70.27	17.00		63.97	20.51		67.36	18.82
	Total		**63.34**	**20.15**		**58.96**	**18.35**		**61.02**	**19.30**

pattern of results. When further splitting the group between those in agreement (43% of males and 36% of females) and those in disagreement (61% of the overall sample) with a stereotype in statistics, the same effects of gender ("disagree": $F(1, 113) = 5.37$, $MSE = .016$, $p = .022$, $\eta^2_p = .045$; "agree": $F(1, 69) = 13.69$, $MSE = .015$, $p < .001$, $\eta^2_p = .166$) and experience ("disagree": $F(2, 113) = 4.28$, $MSE = .013$, $p = .016$, $\eta^2_p = .070$; "agree": $F(2, 69) = 4.01$, $MSE = .013$, $p = .023$, $\eta^2_p = .104$) were present in each group, with the notable absence of an effect of condition in either group.

Confidence

The confidence ratings were collected only once at the very end of the test, by asking participants to report the proportion of problems that they thought they answered correctly. The same 2 (gender) x 2 (condition) x 3 (experience) between-subjects ANOVA was used. Only the main effect of experience was present, $F(2, 194) = 4.43$, $MSE = 360.71$, $p = .013$, $\eta^2_p = .044$, wherein the general level of confidence only increased significantly after having taken two courses in statistics (*Tukey HSD*, $p < .020$). Unlike performance, males and females displayed an equivalent level of confidence, $F(1, 193) = 1.27$, $MSE = 361.72$, $p = .261$, $\eta^2_p = .007$. Again, value affirmation did not influence the scores, $F(1, 193) = .27$, $MSE = 361.72$, $p = .603$, $\eta^2_p = .001$, and did not interact with gender, $F(1, 193) = .415$, $MSE = 361.72$, $p = .520$, $\eta^2_p = .002$). The addition of the level of endorsement of the stereotype in statistics as a covariate (p = .99) did not change this pattern of results. Detailed results of this and other ANCOVAs are reported in Appendix D. For consistency, the sample was split between those in agreement versus those in disagreement with the stereotype. Whereas the "disagree" group showed the same single main effect of experience, $F(2, 113) = 4.43$, $MSE = 325.25$, $p = .014$, $\eta^2_p = .073$, the "agree" group showed no effect at all.

Finally, the calibration of males [$r(95) = .42$, $p < .001$] was compared to that of females [$r(107) = .30$, $p = .001$], but, as in Studies 2 and 3, the difference found in Study 1 was not replicated ($z = .92$, $p = .36$). A scatterplot of those correlations is presented in Figure 8. All other correlations are available in Table 10.

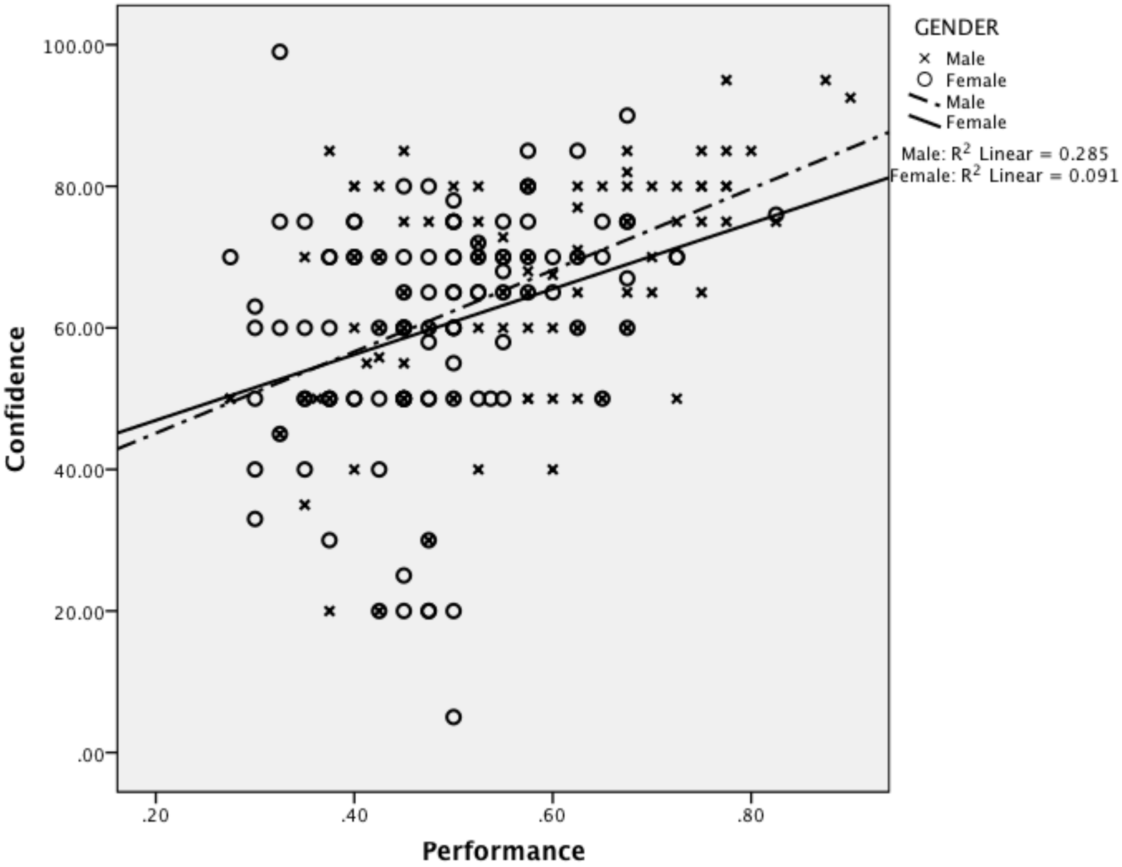


Figure 8. Scatterplot of association between performance and confidence, by gender. Study 4.

Table 10

Correlation Matrix – Study 4

Subscale	**Mean (SD)**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1. CAOS	.52 (.13)	--	.38**	.36**	.57**	-.04	-.06	.25**	.22**	.24**	-.07
2. Conf	61.02 (19.30)		--	24**	.29**	.02	.05	.19**	.25**	.18*	-.01
3. PNI	46.36 (9.97)			--	.45**	.01	.03	.38**	.19**	.29**	-.06
4. CRT	1.43 (1.15)				--	-.01	-.04	.28**	.13	.13	-.22**
5. STstats	2.83 (1.68)					--	.92**	.03	-.09	-.08	-.12
6. STmath	2.87 (1.77)						--	.06	-.04	-.07	-.10
7. Area	3.20 (1.03)							--	.20**	.17*	-.12
8. Formal	2.37 (.82)								--	.54**	-.04
9. Informal	2.98 (.84)									--	.02
10. Word Count	48.84 (33.19)										--

Note. * *p* < .05. ** *p* < .01.

Discussion

Some trends are becoming clear with the addition of this third study examining the potential role of stereotype threat in statistical reasoning. First, there is an inescapable discrepancy in the performance of males and females in statistics – a difference that is not easily explained away or dealt with. Although experience is clearly beneficial for improvement, it does not succeed in closing the gender gap. The hope was that value affirmation, a simple self-focused affirmation technique that has been shown to help reduce performance gaps in some male-dominated areas such as mathematics and physics (e.g., Martens et al., 2006; Miyake et al., 2010), would also help close the gap in statistics. Unfortunately, in three different studies, with various samples, sample sizes, and measures, no effect of value affirmation was found.

The stereotype threat literature also mentions that those who explicitly state their belief in the stereotype are more likely to benefit from any intervention such as value affirmation (e.g., Steele, 1997; Miyake, 2010; Martens et al., 2006), despite the fact that one doesn't necessarily need to believe the stereotype to be affected by it (Steele, 1997). To ascertain that the lack of effect was not due to the inclusion of people without an explicit belief in the stereotype, the analysis was limited to those in agreement with it. Contrary to our expectations, even that subgroup showed no improvement in the value affirmation condition. Overall, we are left with evidence against the existence of a stereotype threat in statistics.

This overall finding is consistent, however, with a recent meta-analysis on the topic of stereotype threat in mathematics (Stoet, & Geary, 2012). Examining the results of replications of the original paper on stereotype threat in mathematics by Spencer, Steele and Quinn (1999), Stoet and Geary found that no more than 55% of those attempts were successful at replicating the original results. In fact, when restricting their sample to the unconfounded studies, that percentage dropped to 30%.

Given these observations, the authors remind researchers to keep deploying research efforts to understand the gender gap on other potentially fruitful avenues as well. For that reason, and given the results of Studies 2, 3 and 4, the next studies will focus on a different factor that purposely targets the cognitive factor of the structural equation model. Specifically, Studies 5 and 6 will focus on mental representations using a novel problem classification paradigm.

STUDY 5

The goal of this fifth study is to examine further the root cause and manifestations of the gender gap in statistics. As seen so far, statistical proficiency does not seem to grace males and females equally, even when prior experience, preparation and grades are the same (Tempelaar, 2006), a finding that is also found in mathematics (Byrnes & Takahira, 1993). Study 2 replicated this finding. Despite equal grades on the prerequisite course, this equivalent preparation nonetheless yielded different scores on the statistical reasoning test. As mentioned by Garfield and Gal (2007), the SRA and the CAOS are designed to probe understanding of statistical concepts. In that sense, they are not surprised when the scores obtained by students on those tests are much lower than the grades obtained on courses that very often emphasize the mastery of computational skills. This implies that, to truly understand the root cause of the gender difference in statistical reasoning, it may be necessary to focus on conceptual understanding rather than on computation.

This is similar to what Quinn & Spencer (2001) suggested in math. In their study, they used mathematical word problems. The task involved identifying the strategy one should use to solve each problem. However, participants did not have to solve the problems. By focusing on the ability of participants to formulate an appropriate solution strategy, which requires identifying the nature of the problems, the researchers eliminated the computational aspects of the solution from being cited as alternative explanations to any gender difference. In addition, participants were later presented with the same problems, with one crucial difference. The problem was now presented in the proper numerical format, thus eliminating the requirement to find the appropriate strategy. Interestingly, where women performed just as well as men when asked to solve the numerical versions of the problems, they were found to underperform in finding a proper strategy to the equivalent word problems in comparison to men. Their performance was also lower on the word problems version than on the numerical problems version. Given those results, the authors concluded that females have more difficulty than males deciphering the nature of mathematical word problems and planning a strategy to solve those problems.

Similarly, the difficulty in identifying the category to which a problem belongs is a great struggle encountered by students in statistics classes (Quilici & Mayer, 2002). Yet, this ability is crucial to the successful solving of statistical problems. Without a clear understanding of the structural features to which one should attend, one is doomed to be unable to compute the appropriate solution before even beginning any calculation. As identifying the type of problem is just as important in statistics as it is in mathematics, the following study will concentrate on the mental representations that participants hold of word problems in statistics, and on their ability to recognize important features in statistical problems.

Quilici and Mayer (2002) argue that structural awareness – the ability to recognize shared structural features that indicate that problems should be solved by the same method – is a necessary quality to reach proficiency in statistics. Structural features are those characteristics important to properly understand and solve a problem, such as underlying principles in physics, underlying problem category in mathematics, or appropriate procedures to solve a problem in statistics. In contrast, surface features are those characteristics that are salient but not necessary to the solution, such as the story line or the shape of the objects depicted on the page. The ability to recognize and use the structural features of problems, both in categorizing and in solving problems, is a sign of expertise (Chi, Feltovich, Glaser, 1981). New standards in the statistics curriculum established by College GAISE recommendations (ASA, 2005) also emphasize the necessity to stress conceptual understanding over the mere knowledge of procedures. In their review, Chi and Roscoe (2002) similarly emphasize the importance of conceptual understanding for expertise.

However, between the novice and the expert, a wide gap exists and the acquisition of this structural awareness can be a lengthy process. For instance, Chi, Feltovich, and Glaser (1981) examined this question in physics. To understand the differences in thinking between novices and experts, participants were asked to categorize 24 physics problems based on how they should be solved. Novices were found to rely heavily on the surface features of the problems. Basically, characteristics of the shapes in the accompanying diagrams (e.g., circular surface, inclined plane) were often offered as the reason behind the classifications made by novices. In contrast, experts were sensitive to the deep, structural features of the problems, i.e., the physics principles that were at the heart of each problem, regardless of the shapes present in the diagrams. Thus, it was claimed that experts, unlike novices, are able to look past the superficial surface features of problems, focusing instead on the deep or structural features that represent principles when thinking about how they would solve the problems.

The same holds true for statistics. For example, Rabinowitz and Hogan (2008) used a triad task requiring participants to select the best match to a target problem. For each target problem, participants could select between problems that were similar to the target problem either due to the surface features of the storyline, or due to the underlying statistical characteristics. They found that even participants with extensive statistics training, i.e., those with over 4 courses in statistics, matched some of the problems based on their surface characteristics, although less often than those with less experience. This finding was also replicated with teachers of various experience levels (Hogan & Rabinowitz, 2009).

Quillici and Mayer (1996) also found that structural awareness could be fostered through direct instruction. Participants were exposed to examples of problems that could be solved by calculating a t-test, a correlation, or a chi-square test. However, where some participants saw examples that made salient the tests to use (the structure-emphasizing condition), others saw examples that made salient the theme (e.g., weather) of the problem (the surface-emphasizing condition). After training, participants were asked to categorize 12 problems. Crucially, those problems could be sorted either by one of the three tests (e.g., classifying together all problems that can be solved with a correlation) or by one of the four themes e.g., classifying together all problems that have a weather- related storyline). Those exposed to structure-emphasizing examples sorted the problems based on structure more often than students trained with surface-emphasizing examples, and also showed greater proficiency in choosing the appropriate test to actually solve problems in a subsequent experiment. However, even after some specific short-term training sessions, students were still having difficulty distinguishing between t-test versus chi-square problems. The same results were obtained when experience was measured (categorizing based on how many statistics courses participants had taken) rather than manipulated (providing short training sessions) (Quilici & Mayer, 2002).

One limitation to the interpretation of the Quilici and Mayer's (2002) results is that, unlike Chi, Feltovich, and Glaser (1981), participants' rationales for their classifications were not collected. Thus, it is uncertain whether poor performance was due to a lack of knowledge or simply due to inattention errors. Collecting rationales behind performance would have allowed a deeper, qualitative analysis of the data. Therefore, in the current study of this dissertation, a combination of Chi et al.'s (1981) framework and Quilici and Mayer's (2002) methodology is used, with some modifications. For instance, Chi et al.'s (1981) paradigm to test the development of expertise and its related conceptual understanding is used, though extending it to the statistical realm. Then, unlike prior research by Quilici and Mayer (1996; 2002), data characteristics (i.e., type of data and number of variables) rather than just different types of tests (i.e., t-test, correlation, chi-square) are used to serve as deep, structural features. However, consistent with statistical decision trees found in textbooks and online, the combination of both data characteristics (type of data and number of variables) allows the identification of what statistical test should be used to analyze the data at hand (e.g., whether to use a one-way ANOVA, two-way ANOVA, chi-square test of goodness of fit, chi-square test of independence). In that sense, classifying over only one characteristic, in comparison to classifying each problem based on the combination of the two characteristics, can be seen as representing an intermediate level of statistical sophistication.

Multiple authors have referred to intermediate levels of sophistication in studying statistical reasoning. For example, in coding the responses of participants to scenarios involving the law of large numbers, Fong, Krantz, and Nisbett (1986) used three levels: Non-statistical, poor statistical reasoning, and good statistical reasoning. A response coded as "poor statistical" was one that did not focus on the surface features of the problem, but did not provide the correct statistical explanation either. Similarly, Lavigne, Salkind, and Yan (2008) also used multiple levels in coding participants' rationales. In addition to the surface-based (focused on the storyline or context) and the principled (organized around principles or solution methods), a pre-structural representation category was created. This category allowed the researchers to better capture the variation in rationales, and resonates with the idea that learning keeps building on previous conceptions (Biggs & Collis, 1982). This "principled problem representation in the making" includes abstraction of problem features that help the decision process, as when using a statistical decision tree for instance. Their final coding scheme actually included five levels of sophistication, with three of them related to the pre-structural level of understanding.

Halfway between the three levels of Fong et al. (1986) and the five levels of Lavigne et al. (2008), Shaughnessy (1992) propose four levels of statistical sophistication in his cognitive development model. The first level achieved, *non-statistical*, is equivalent to the surface-based reasoning discussed earlier. Those in the second level, naive-statistical, are still influenced by salient characteristics of the data and their reasoning does not demonstrate a deep understanding of the statistical ideas. In line with the inclusion of experience as a variable in this study, Shaughnessy sees the upper two levels as resulting from formal instruction in statistics. Thus, only those with previous experience in statistics should be able to display these levels of statistical reasoning. Those at the emergent-statistical level will display a higher level of understanding of statistical ideas. As with Lavigne et al.'s (2008) upper levels of pre-structural understanding, multiple characteristics of the data should be recognized and integrated. Finally, Shaughnessy notes that very few people will ever reach the pragmatic-statistical level of sophistication. This level is comparable to the principled understanding characteristic of experts (Chi et al., 1981; Silver, 1981).

With these previous findings in mind, it was decided in the current study to combine the categorization paradigm used in Chi et al (1981) with the use of statistical word problems. The problems that participants were asked to categorize included both statistical characteristics (i.e., type of data, number of variables) reflecting the deep structural features, and content themes (akin to the weather problems) reflecting surface features. In that sense, novices should be more likely to create sets corresponding to the content themes; intermediates should be more likely to create sets corresponding to either one of the data characteristics; and those with more experience should be more likely to create sets corresponding to the test appropriate when taking both types of data characteristics into consideration (e.g., One-way ANOVA). The task also required participants to explain the reasoning behind each of their classifications. Logistically, problems were presented individually on cards (as done by Chi et al., 1981) rather than all on a same sheet (as done by Quilici & Mayer, 2002). Indeed, research in the area of embodied cognition suggests that physically manipulating stimuli, especially varying the distance between them, may help performance when separate categories are involved (e.g., Lakens, Schneider, Jostmann, & Schubert, 2011). Thus, using cards to physically sort the problems into categories should help participants better discriminate among them and improve their performance.

Overall, it was predicted that males would perform better than females, with good performance defined as classifying problems more accurately (i.e., sorting based on deep, structural features) and as providing rationales reflecting an awareness of the deep features of the problems. In addition, it was expected that increased experience would lead to better performance on both measures, and (based on the findings from Studies 1 to 4) that no interaction would be found between experience and gender.

Method

Participants

Undergraduate students from the University of Waterloo participated for course credit. Eight students with missing data on any of the two main predictor variables (i.e., gender, number of stats courses taken) and six students who indicated not being fluent in English were eliminated from the analysis. In the end, 219 participants (103 males, 116 females; M_{age} = 19.87*,* *SD* = 2.03 – ranging from 17 to 31) with varying levels of experience (0: N = 91, 1: N = 88, 2^{+}: N = 40) were included in the analysis. Participants came from fields deemed generally non-quantitative (e.g., music, philosophy) to extremely quantitative (e.g., mathematics, statistics) [Generally non-quantitative = 16 (7%), Minimally quantitative = 20 (9%), Moderately quantitative = 65 (30%), Highly quantitative = 104 (47%), Extremely quantitative = 14 (6%)].

Design

Two of the between-subject predictor variables used in the previous studies were relevant for this study, namely experience - operationalized as the number of statistics courses taken in university (i.e., 0, 1, and 2^+) - and gender. Three new dependent variables of interest were assessed: (1) the number of sets created, (2) the number of deviations from the ideal categorization that are committed (see explanation below), as well as (3) the quality of the rationales provided to explain why each set was created.

Materials

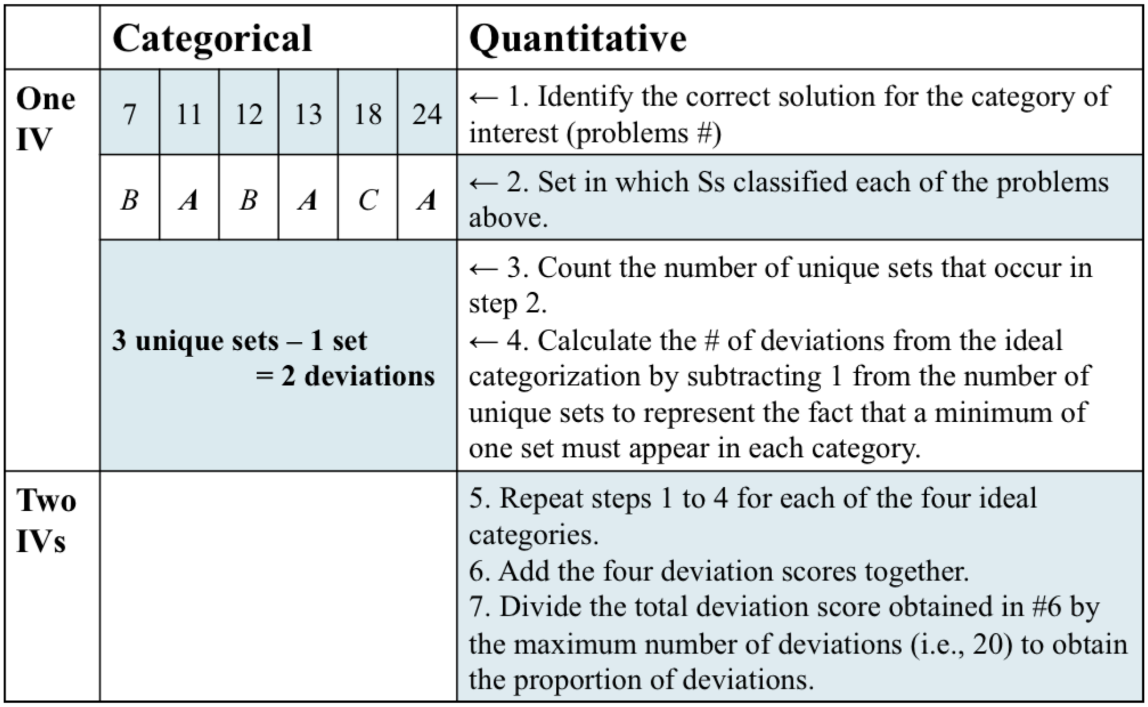
Twenty-four problems selected from an introduction to statistics textbook by Goldman and Weinberg (1985) were modified slightly to fit the purpose of this study. The problems were distributed equally among twelve cells within three different 2 x 2 tables. Each of the three tables is related to one of three main themes (i.e., A- health; B- demographics; C- goods and services), which represent the surface features of the problems. More importantly, each theme contains eight problems that are defined by two structural characteristics: Type of data (i.e., continuous vs. categorical) and number of independent variables (i.e., one variable vs. two variables). Those characteristics represented the deep features of the problems – those that inform the type of statistical test required to analyze the data at hand. Stated differently, each of the four ideal groupings contains six problems. Each problem was randomly assigned a number from 1 to 24 when the problem cards were created. The list of all problems – organized based on the four ideal groupings (1- continuous DV/one IV; 2- continuous DV/two IVs; 3- categorical DV/one

IV; 4- categorical DV/two IVs) – is available in Appendix F. To help facilitate comprehension, it should be noted that *set* refers to "a group of problems created by the participant"; *classification* refers to "the overall distribution of problems into sets by the participant"; *category* refers to "the feature of interest for the analysis, whether superficial or structural"; *grouping* refers to "a specific group of problems determined by the category of interest"; and *categorization* refers to "the ideal overall distribution of problems as determined by the category of interest."

The task instructions were consistent with the task instructions used by Chi et al. (1981). They were strikingly simple and open to interpretation, only specifying that the goal was to classify the problems based on how the participant would solve them (see Appendix F). The sets created by participants were reported in an arbitrary order, and identified in alphabetical order (e.g., A, B, C, D, E…). The content of each set was determined and reported by the participant on their answer sheet. The task structure thus allowed the determination of two quantitative measures as proxies for statistical sophistication, namely (1) the number of sets created and (2) the number of deviations from the ideal categorization (to be explained shortly).

The number of sets created is of interest given that the instructions do not specify how many sets should be created. This measure involves counting the number of sets reported by participants on their answer sheet. Simply, if a participant distributed the 24 problems among six sets, her score for the number of sets would be 6. If the problems are distributed among three sets, the score is 3.

For the calculation of the number of deviations from the ideal categorization, the steps are described below and illustrated in Figure 9. Essentially, they involve comparing the sets created by participants to the ideal four groupings solution (1- continuous DV/one IV; 2- continuous DV/two IVs; 3- categorical DV/one IV; 4- categorical DV/two IVs). Each ideal



	Categorical						Quantitative
One IV	7	11	12	13	18	24	← 1. Identify the correct solution for the category of interest (problems #)
	B	*A*	*B*	*A*	*C*	*A*	← 2. Set in which Ss classified each of the problems above.
	3 unique sets – 1 set = 2 deviations						← 3. Count the number of unique sets that occur in step 2. ← 4. Calculate the # of deviations from the ideal categorization by subtracting 1 from the number of unique sets to represent the fact that a minimum of one set must appear in each category.
Two IVs							5. Repeat steps 1 to 4 for each of the four ideal categories. 6. Add the four deviation scores together. 7. Divide the total deviation score obtained in #6 by the maximum number of deviations (i.e., 20) to obtain the proportion of deviations.

Figure 9. Calculation process for deviation scores.

grouping contains six of the 24 problems. Categorizing the problems based on the combination of the two types of deep structural features in the problems thus yields four groupings: Grouping A includes problems #5, 10, 20, 21, 22, 23; grouping B includes problems #1, 4, 8, 16, 17, 19; grouping C includes problems #7, 11, 12, 13, 18, 24; and grouping D includes problems #2, 3, 6, 9, 14, 15. These groupings then serve as the basis to calculate the deviation score.

Concretely, suppose that you are comparing the participant's classification in sets to the ideal grouping C above, which contains the following six problems: #7, 11, 12, 13, 18, 24. First, you need to know whether the participant has classified these problems together or not. To do so, the letter of the set in which each of those six problems appears in her classification is noted. For example, problems #11-13-24 could appear in set A, problems #7 and 12 appear in set B, and problem #18 appear in set C. For any participant, the problems of interest will always appear either in one same set (e.g., all in set A), or in different sets (e.g., some in set A, some in set B, and some in set C). Counting the number of sets in which the six problems of interest are classified constitutes the next step. In the example above, the participant has classified the six problems of interest into three sets. Thus, the participant receives a classification score of 3, indicating that three different sets had been used in classifying the problems from this ideal grouping. If all six problems had been classified in set A, the participant would have received a classification score of 1, indicating a perfect classification. On the other hand, if each of the six problems of interest had appeared in different sets, the participant would have received a classification score of 6. The same process is done for each of the remaining ideal groupings. The number of classifications for each grouping could range from 1 (perfect classification) to 6 (poorest classification). Finally, to obtain a deviation score, 1 was subtracted from the classification score for that grouping – to represent the fact that a minimum of one set must obligatorily appear in each grouping. Therefore, when dealing with four ideal sets, the total number of misclassifications can range from 0 to 20, which is equivalent to an average of 0 to 5 possible deviations per ideal grouping. Lastly, a proportion score is calculated to facilitate interpretation of the measure[5.]

[5] Of course, the same process can be done to compare the number of deviations for any other category of interest (i.e., based on themes, based on type of data, or based on number of independent variables). The only difference is that the possible number of deviations would vary based on the number of problems that the category of interest establishes for each grouping. For instance, given that two groupings of twelve problems are ideal when based on the type of data, the total number of deviations can vary from 0 to 22, which is equivalent to an average of 0 to 11 deviations for each of the two groupings in that category. For the category based on the surface features (i.e., themes), this represents three groupings of eight problems. Thus, the number of deviations can range between 0 and 21, which is equivalent to an average of 0 to 7 deviations for each of the three groupings in that category. Given the variable number of deviations based on the category of interest, a proportion score can be calculated for each category to allow a comparison of the deviation scores across categories. This proportion is calculated by dividing the number of deviations committed in a specific category by the highest possible number of deviations in that category. It is important to note that, whereas a low proportion of deviations is desirable when dealing with structural features, a high proportion of deviations is preferable when dealing with surface features.

Another important aspect of the participant's task involves providing a rationale for each set created. These qualitative data allow us to prevent spurious classifications – either correct or incorrect – from biasing the results by affecting the quantitative proxies. For instance, it can be uncertain whether poor performance on the classification task was due to a lack of knowledge or simply due to inattention errors. Thus, the rationales provided by participants with each of their sets provided an additional proxy for statistical sophistication. Research assistants – blind to the gender and experience level of the participants – were trained to code the quality of those rationales. The coding proceeded as follows. First, an initial coding scheme was established (see Table 11). Two research assistants were then briefed on the different categories and coded the entire dataset once. Difficulties with the initial coding scheme were discussed and the scale was modified to improve the precision of the coding. A third research assistant joined the team of coders for the second round of coding. The RAs first coded approximately 300 items each with the new coding scheme. As no substantial problem was encountered with the new scale, the full set of data was coded by the three RAs. At the end of this round, the inter-rater reliability (average intra-class correlation coefficient) was .95. Given the very good agreement, focus was turned to mismatches in coding. RAs were given the list of rationales for which there was disagreement and asked to reconsider the coding of those items. Any remaining mismatched ratings were discussed between the three coders and the main researcher until perfect agreement was met.

The original coding scheme was then modified to match Shaughnessy's (1992) framework for statistical sophistication (see Table 11). Shaughnessy's (1992) framework was used for three reasons. Firstly, each level of that 4-point scale corresponds squarely to two levels of the initial 8-point scale (see Table 11). Secondly, this rating system produced homogeneous variance across the groups, thus circumventing the problem of heterogeneity of variance present when using the scores of the 8-point scale. Thirdly, this scale provided a stronger theoretical grounding for the results. Using this new scoring system, each rationale provided by participants received a score. To represent simply the collection of rationales provided by each participant, the modal score for each participant was calculated and used in all subsequent analyses. The mode was chosen as the most stable representation of the ability of the participants, and as a way to protect against potential coding errors, especially when the number of sets created was low. If multiple modes were present, the mean value of those modes was calculated and used for analysis.

Table 11

Coding Scheme

Sophistication Level	Reasoning included in each level (original scheme)
(1) Non-statistical	No reason given. / Say that it does not fit with other problems.
	Focus on theme in the problem.
(2) Naïve-statistical	Focus on type of conclusion that could be drawn, without clear statistical consideration.
	Noticing the types of data reported (e.g., means). -OR- Noticing the breakdown of variables into levels.
(3) Emergent-statistical	Recognizing that the solution bears an association of some kind between the variables. Inappropriate test mentioned.
	Recognizing the type of data reported and the need to compare the various groups.
(4) Pragmatic-statistical	Stating a statistical test appropriate for the type of data.
	Stating a statistical test appropriate for the type of data, while acknowledging the number of variables as influencing the choice of test.

Note. Based on Shaughnessy's (1992) framework for statistical sophistication.

Procedure

The study took place in a lab and participants were tested in groups of 1 to 3. Participants read the information letter and signed the consent form prior to completing the tasks. First, a set of 24 problems, each appearing on a separate card, was distributed to participants. The task instructions asked participants to group the problems based on how they would solve them, as per Chi, Feltovich, & Glaser (1981) (see Appendix F). Importantly, participants did not have to solve the problems. Second, once their classification was completed, participants were asked to write down the problem numbers included in each of their sets and to provide a brief written explanation of the basis on which they created each set. Once the main task was done, participants were asked to complete a short demographic questionnaire based on Schield (2005). Finally, participants were given the opportunity to ask questions about the study and received a feedback letter.

Results

In this study, two independent variables were of interest: gender and experience in statistics. As noted above, the analyses centred around three different dependent variables: number of sets created, number of deviations from the ideal categorization (i.e., based on deep, structural features), and quality of rationales provided. It was expected that females' classifications would reveal a greater number of deviations, and that the quality of their rationales would be lower than that of males. It was also expected that the number of deviations would decrease with increased experience, and that the quality of rationales would be higher with increased experience. The number of sets created is a relevant dependent variable as it may influence the other, more conceptually relevant

Table 12

Descriptive Statistics – Study 5 – Deviations from Ideal & Quality of Rationales

			Male			**Female**			**Total**	
	# Stats courses taken	n	*Mean*	*SD*	n	*Mean*	*SD*	n	*Mean*	*SD*
Number of Sets	0	**45**	5.62	2.67	**46**	5.74	2.33	**91**	5.68	2.49
	1	**37**	4.73	1.63	**51**	5.86	2.80	**88**	5.39	2.43
	2^+	**21**	4.90	1.64	**19**	5.47	2.57	**40**	5.17	2.12
	Total	**103**	5.16	2.17	**116**	5.75	2.57	**219**	5.47	2.40
Deviations	0		.47	.25		.49	.28		.48	.26
	1		.36	.20		.48	.30		.43	.27
	2^+		.40	.22		.45	.29		.42	.25
	Total		**.41**	**.23**		**.48**	**.29**		**.45**	**.26**
Rationales	0		1.58	.65		1.48	.55		1.53	.60
	1		1.68	.63		1.65	.66		1.66	.64
	2^+		2.05	.86		1.68	.67		1.88	.79
	Total		**1.71**	**.71**		**1.59**	**.62**		**1.64**	**.66**

DVs. For that reason, it needed to be analysed first, even though no systematic effects were expected. Descriptive statistics for all three dependent variables are presented in Table 12.

Quantitative measures: Number of Sets & Deviations from Ideal

Number of Sets. The task instructions did not specify how many sets were to be created, thus allowing this variable to be analysed. Participants each created between 2 and 12 sets (*M* = 5.47, *SD* = 2.403, *Median* = 5, *Mode* = 4). The highest numbers of sets occurred mostly when participants focused on pure surface features (e.g., “Looked at health problems”, “Looked at family income, unemployment"). Despite males creating slightly fewer sets than females on average (M_m = 5.16, *SD* = 2.17; M_f = 5.75, *SD* = 2.57), the difference did not reach significance, *F*(1, 213) = 3.03, *MSE* = 5.725, *p* = .083, η^2_p = .014. Experience did not influence the number of sets created either, *F*(2, 213) = .834, *MSE* = 5.725, *p* = .436, η^2_p = .008, even though the number of sets did decrease slightly with increased experience. Finally, the two independent variables did not interact, *F*(2, 213) = 1.00, *MSE* = 5.725, *p* = .371, η^2_p = .009. Thus, the results did not support the hypothesis that experience would lead to the creation of a different number of sets.

Number of Deviations from Ideal. From the sets' composition (i.e., what problems a participant placed in any given set s/he created), a measure of deviation from the ideal categorization was computed. The specifics of the scoring process are explained above in the method section and illustrated in Figure 8. The expectation was that increased experience would lead to a decreased number of deviations. A 2 (gender) x 3 (experience) between-subjects ANOVA was used to analyze these data. Surprisingly, no main effects were found, even though the means were in the expected direction, with females' classifications deviating more from ideal than males' classifications, *F*(1, 213) = 2.90, *MSE* = .069, *p* = .090, η^2_p = .013., and with deviations decreasing with experience, *F*(2, 213) = 1.228, *MSE* = .069, *p* = .295, η^2_p = .011. Gender and experience also did not interact, *F*(2, 213) = .77, *MSE* = .069, *p* = .463, η^2_p = .007 .

To address the possibility that high levels of deviations may only be reflective of how many sets one created, a second analysis of variance was conducted, this time using the number of sets created as a covariate. Even though number of sets emerged as a significant covariate ($p <$ .001), this finding did not influence the previous results.

So far, this focus on quantitative indicators provides limited insight into variations in mental representations and statistical sophistication. Thus, to further examine mental representations in statistical reasoning, a qualitative indicator was also analysed, as it is well known that using multiple measures can increase the richness of our understanding of a phenomenon (e.g., Johnson & Onwuegbuzie, 2004).

Qualitative Measure

Quality of Rationales. A main goal of this study is to examine the impact of experience on the quality of statistical reasoning. Even more so, it is hoped that it can help shed light on the underlying cause of the gender gap. To obtain a qualitative picture of individuals' mental representations in statistics, participants were required to explain *why* they had created each set, i.e., *why they saw those problems as fitting together*. As discussed above, the scores were obtained through rigorous rounds of coding. A 2 (gender) x 3 (experience) between-subjects ANOVA was used to analyse those scores. Although the quality of the rationales provided by males and females was more similar than expected, $F(1, 213) = 2.97$, $MSE = .426$, $p = .086$, $\eta^2_p = .014$, a main effect of experience was present, $F(2, 213) = 3.77$, $MSE = .426$, $p = .025$, $\eta^2_p = .034$, with the quality of rationales increasing with experience (*Tukey HSD*, $p = .015$) after two courses in statistics, though only in comparison to those with no experience ($M_0 = 1.53$, $SD = .60$; $M_1 = 1.66$, $SD = .64$; $M_{2+} = 1.88$, $SD = .79$). Even when using the more liberal *Fisher's LSD* instead of *Tukey's HSD*, the comparison of those with one versus those with at least two courses in statistics remained non-significant ($p = .084$).

Despite the lack of an interaction between gender and experience, $F(2, 213) = .915$, $MSE = .426$, $p = .402$, $\eta^2_p = .009$, it felt prudent to nonetheless explore the possibility that each gender responds differently to experience, and to verify at what point in the process the quality of rationales becomes increasingly and significantly better. For each gender, a one-way ANOVA was used to analyse the impact of experience. Interestingly, whereas males appeared to improve the quality of their rationales with added experience, *F*(2, 100) = 3.39, *MSE* = .475, *p* = .038, females' quality of rationales appeared to stagnate, *F*(2, 113) = 1.19, *MSE* = .475, *p* = .309. Indeed, using *Fisher's Least Significant Difference* test for multiple comparisons, the reasoning of males with the most experience significantly differed both from the reasoning of those with no experience (*p* = .011) and of those with the experience of only one stats course (*p* = .051), which was not the case for females with the most experience when compared to either females with no (p = .225) or with only one (p = .824) stats course.

Discussion

In this study, both quantitative and qualitative measures were used to index participants' level of statistical sophistication. The inclusion of coded rationales as a measure proved particularly useful. With added experience, participants explained their classifications with increased reference to statistical features contained in the problems. Further investigation of this effect demonstrated that it is males, but not females, who show a change in the quality of their rationales with added experience. This lack of improvement for females in the ability to detect and use structural features in statistical problems could be one clue toward the explanation of the gender gap in statistics.

On the other hand, quantitative indicators were not as informative. Despite some trends in mean differences, those differences were not statistically significant. The freedom participants had to create as many sets as they wanted could have negatively impacted their propensity to generate sets based on deep structural features. Rather, the lack of clear guidelines may have encouraged them to rely on the most salient solution. Indeed, surface rather than structural features are naturally more salient for non-experts (Lavigne, Salkind, & Yan, 2008). Alternatively, the findings could truly represent participants' level of knowledge and their lack of integration of statistical entities into functional representations in memory. However, to make sure that participants did not rely on surface features simply due to a lack of understanding of the instructions or due to a tendency to settle for the most salient solution, a follow up study was included with a critical change: an additional task constraint to control the number of sets created.

Lavigne, Salkind, and Yan (2008) were the first to examine the impact of varying task format on mental representations in statistics, asking whether one type of representation is more likely to be elicited when the instructions are modified. However, their findings – that a more constrained or specific task format helps the identification of structural features – were based on a sample of three participants. In the next study (Study 6), the task will be constrained, as participants will be instructed to create exactly four sets. Where some may argue that this constraint may lead to fewer deviations from the ideal, this is an empirical question. Furthermore, as demonstrated in math (Quinn & Spencer, 2001), females' difficulty with establishing a solution strategy may be the root cause of their general difficulties in mathematics. In their study, females showed the same ability as males in solving numerical math problems, a task with well-defined constraints. Yet, when asked to plan how they would solve some math word problems, which were a more complex and less constrained but equivalent form of the numerical problems mentioned earlier, the performance of females dropped. In contrast, the performance of males was comparable across the two tasks. Thus, it is possible that constraining the task will especially benefit the performance of women and leave males' performance intact.

Another concern in this study was the unequal cell sizes due to the difficulty of finding participants with higher levels of experience. In the following study, one goal was to achieve a minimum of 25-30 participants per cell. To achieve that goal in a timely manner, graduate students – as in Study 1 – were also recruited.

STUDY 6

In this study, we address the possibility that an overly unconstrained task format may have impeded the performance of participants on the task. Here, the same procedure as for Study 5 was used, except for the fact that participants were instructed to classify the problems in exactly four categories. Based on Lavigne, Salkind, and Yan (2008), we know that different instructions can lead to different mental representations in the domain of statistics. However, their study used a meagre sample of three males. This finding is thus in dire need of replication and extension. Nonetheless, a change in task format could potentially have a very positive impact on women. As demonstrated by Quinn and Spencer (2001), when women are shown clearly defined mathematics problems, they can solve them just as well as men. The authors concluded that women have difficulty identifying the type of problem and the appropriate course of action when faced with word problems, but that they can solve the problem just as well as men when the problem category is well defined, such as when a mathematical word problem has been translated into its numerical form. If women also struggle with identifying the appropriate strategy when presented with statistics word problems, then constraining the task should help females improve their performance on the classification task.

With these new instructions, the number of sets created is constant and no longer needs to be analyzed. Again, it is expected that increased experience will lead to more accurate classifications by participants (i.e., based on deep, structural features). In addition, the rationales provided by participants should reflect their awareness of the deep features of the problems, and become increasingly sophisticated with experience. In line with studies 1-4, and speculating that the open instructions may have blurred the results from study 5, we also expected that males would perform better than females on both measures, but that the change in task format might be particularly beneficial for females' performance.

Method

Participants

Undergraduate and graduate students from the University of Waterloo participated in this study for course credit or remuneration. Graduate students were recruited in this study due to the difficulty of signing up participants having taken two courses in statistics in Study 5. This ensured that enough participants were included in each cell for the analyses. All participants having taken at least two courses in statistics were included in the same level of experience (i.e., 2^+). Participants with missing data on any of the two main predictor variables (i.e., gender, number of stats courses taken), participants who indicated not being fluent in English, and participants who indicated having taken more than 10 courses in statistics were eliminated from the analysis. In the end, 208 participants (101 males and 107 females; 0: N = 83, 1: N = 68, 2^+: N = 57; M_{age} = 21.27, *SD* = 3.16, ranging from 18 to 35) were included in the analysis. The distribution of participants across areas was comparable to the previous studies [Generally non-quantitative = 10 (5%), Minimally quantitative = 28 (14%), Moderately quantitative = 69 (33%), Highly quantitative = 88 (43%), Extremely quantitative = 12 (6%)].

Design and Materials

Apart from the elimination of number of sets from the list of dependent variables, the design was the same as in the previous study, with gender and experience as independent variables, and deviations and rationales as dependent variables. As with Study 5, a deviation score of 0 represents a perfect classification of the problems, whereas a deviation score of 1 represents a total departure from the ideal categorization. For the rationales, the same procedure and coding scheme as Study 5 are used. The scores thus range from 1 (non-statistical) to 4 (pragmatic-statistical) (see Table 11).

The answer sheet was modified to reflect the need to create a fixed number of sets, and the instructions were reworded slightly to make sure that participants understood the goal of classifying the problems based on how they should be solved (see Appendix F). To test the influence of this instructional manipulation, data from the two studies were combined in a final series of analyses.

Procedure

Excluding the new limitation to the number of sets that one could create, the same procedure as in Study 5 was used.

Results

For this second study addressing mental representations, gender and experience remained as independent variables of interest, with deviations from ideal categorization and quality of rationales used as dependent variables. Given the change in task format, the results could potentially diverge from the ones found in Study 5. This change in task format was likely to increase the salience of ideal categorization and to yield better performance. In particular, the number of deviations was expected to decrease and the quality of the rationales provided expected to increase. Results for Study 6 will be presented first, followed by the analyses testing the impact of constraining the task format across the two studies. When task format is analyzed, the sample size for the combined studies increases to N = 427, thus influencing the degrees of freedom reported. Descriptive statistics for Study 6 are presented in Table 13.

Table 13

Descriptive Statistics – Study 6 – Deviations from Ideal & Quality of Rationales

		Male			**Female**			**Total**		
	# Stats courses taken	n	*Mean*	*SD*	n	*Mean*	*SD*	n	*Mean*	*SD*
Deviations	0	**43**	.54	.27	**40**	.45	.27	**83**	.50	.27
	1	**33**	.45	.26	**35**	.43	.25	**68**	.44	.25
	2^+	**25**	.41	.24	**32**	.45	.25	**57**	.44	.24
	Total	**101**	**.48**	**.26**	**107**	**.45**	**.25**	**208**	**.46**	**.26**
Rationales	0		1.83	.64		1.98	.59		1.98	.59
	1		1.88	.64		2.04	.62		2.04	.62
	2^+		1.72	.71		2.13	.79		2.13	.79
	Total		**1.82**	**.65**		**2.04**	**.66**		**2.04**	**.66**

Quantitative Measure: Deviations from Ideal

As with the previous study, a 2 (gender) x 3 (experience) between-subjects ANOVA was conducted, with the expectation that males would outperform females and that the number of deviations would decrease with experience. In the current study, neither gender, *F*(1, 202) = .36, *MSE* = .066, *p* = .550, η^2_p = .002, nor experience, *F*(2, 202) = 1.41, *MSE* = .066, *p* = .248, η^2_p = .014, had an effect on the proportion of deviations from ideal. Their interaction was not significant either, *F*(2, 202) = 1.09, *MSE* = .066, *p* = .340, η^2_p = .011. This pattern is similar to what was found in Study 5, which the next analysis confirms. The new analysis of interest examines whether the constrained task format, now requiring participants to create four and only four sets, helps improve performance. Unfortunately, despite the more restrictive set of instructions, participants' classifications deviated from the ideal categorization just as much as in the previous study, $F(1, 415) = .45$, $MSE = .067$, $p = .501$, $\eta^2_p = .001$. Adding covariates to the analyses did not change the overall patterns of results (see Appendix D).

Qualitative Measure: Quality of Rationales

To establish the role of gender and experience in statistical sophistication, a 2 (gender) x 3 (experience) between-subjects ANOVA was conducted on the rationales given with each set. Again, based on Shaughnessy's (1992) framework, the scores range from 1 ("non-statistical") to 4 ("pragmatic-statistical") (see Table 11). Except for the non-significant interaction, the pattern of results was quite different from that observed in Study 5. This is confirmed with a main effect of task format when it is added as a factor in the analysis, $F(1, 415) = 13.29$, $MSE = .439$, $p < .001$, $\eta^2_p = .031$. First, unlike Study 5, experience did not lead to better performance, $F(2, 202) = .16$, $MSE = .439$, $p = .486$, $\eta^2_p = .007$. Second, despite the presence of a significant gender difference, $F(1, 202) = 6.59$, $MSE = .439$, $p = .011$, $\eta^2_p = .032$, the direction of the effect was contrary to the prediction. Indeed, in this experiment, it was the quality of females' rationales that surpassed the quality of males' rationales. This change of trend is corroborated by a significant interaction of gender with instructions type when both Study 5 and Study 6 are combined within one analysis, $F(1, 415) = 9.17$, $MSE = .432$, $p = .003$, $\eta^2_p = .022$. In other words, where males' performance did not vary with the constrained task format, $t(202) = 1.13$, $SE = .10$, $p = .258$, females' performance benefited significantly from an increasingly constrained task format, $t(221) = 5.31$, $SE = .09$, $p < .001$.

Discussion

Once again, both quantitative and qualitative indicators were used to measure performance on this classification task. As in Study 5, using a quantitative measure did not allow the distinction of performance across genders and experience levels, whereas using the qualitative measure revealed significant variations in performance. The most important difference occurred due to the change in task format. Indeed, females not only improved upon their previous performance, but also performed significantly better than males in this study when the task format was constrained. In contrast, males' performance on the task remained unchanged despite the more constrained task format.

It is important to note that the differences in findings across the studies cannot be due to coding. For both studies, coders were blind to the gender and experience of the participants. As such, this difference could not have been created due to an expectation bias. Rather, this finding is congruent with the cognitive approach used by Byrnes and Takahira (1993) in their examination of gender differences on SAT-math items. This cognitive approach identifies the ability to define a problem as an important component of skilled performance and stipulates that gender differences exist because males perform certain cognitive operations more effectively than women. Those operations include deciding on the proper problem solving strategy. This lower ability of women to decide on the proper strategy may make them look more readily for cues to support their choice. Although more research is needed to fully understand the underlying cause of this differential effect of task format on gender, it is possible that males simply did not pay further attention to the additional cue that the constrained task format provided.

Noticeably, experience did not have its expected effect in this study. Even though the means typically varied in the expected direction, the effects were not significant. Shaughnessy (1992) does warn us that progression through the various levels of statistical sophistication is usually slow. Reaching the third level, emergent-statistical, requires a good amount of formal education. As such, only a very small percentage of the population will ever reach the pragmatic-statistical level. Even when they reach a high level of statistical sophistication, consistent use of that knowledge is not guaranteed (Hoffrage et al., 2000; Kahneman et al., 1982). For example, in Fong, Krantz, and Nisbett's (1986) study, when presented with a series of multiple problems, the technical experts were found to think statistically about those problems only 80% of the time. Rabinowitz and Hogan (2008) also found that graduate students revert to using surface features a sizable amount of time to guide them in sorting statistical problems. Perhaps the range of experience used in this study was too limited. For instance, Chi et al. (1981) contrasted the performance of novices to that of experts who were professors in physics. Graduate students were only considered intermediates. Although having access to a greater range of levels of experience could help us better understand the development of statistical sophistication, the sample used in this study is of great interest as it represents more closely the level of experience that today's citizens and knowledge workers are likely to have in general.

Again, modifying the task format from 'unconstrained' in Study 5 to 'constrained' Study 6 had different effects across genders on the quality of the rationales that participants provided, but not on the quantitative measure of deviations from ideal. Specifically, for males, this modification in task format had no impact on their performance. Indeed, the quality of their rationales was at par across the studies. In contrast, constraining the problem boundaries helped females provide better rationales. This is similar to the finding in mathematics where females' performance drops when they are faced with ill-defined word problems rather than simple, well-defined numerical problems (Quinn & Spencer, 2001). This finding can be seen as grim from the standpoint of everyday life where data are often messy, as this may prevent women from performing at their full potential. On the other hand, the identification of such an important aspect for the success of women in statistical reasoning can be positively used as a stepping stone for the development of effective individualized pedagogical interventions.

GENERAL DISCUSSION

The unprecedented amount of data available in today's society is both exciting and challenging. Making sense of these data to inform personal, business, and societal choices requires citizens and decision-makers to have at least some degree of statistical literacy. The new faces of work and access to information have already influenced the structure of education in the field of statistics, but some puzzles remain unsolved, notably that of a gender difference. This underperformance of women in statistics is particularly concerning in light of the finding that low numeracy is detrimental to employability, especially for women (Parsons & Bynner, 1997, 2005).

This interest in statistical reasoning also comes at a time when statistics education is redefining itself, notably as a science of data separate from mathematics. New recommendations in line with the GAISE report (ASA, 2005), a set of guidelines for statistics education, are now being implemented in the classroom. For instance, the use of computer tools to perform computations is encouraged to allow students to focus more on the conceptual understanding of statistics rather than the mechanics of conducting the analysis.

New assessment tools have been created to evaluate conceptual understanding and statistical competence, as defined by the new curriculum. In particular, the SRA and the CAOS are designed to assess a wide range of statistical concepts covered in high school and in introduction to statistics classes at the college level. Both tests have been used in research and are recognized as standardized instruments in the field of statistics education. An advantage of the SRA over the CAOS is how it measures both correct reasoning and misconceptions. Misconceptions are an especially informative metric from an educational standpoint as they are generally considered to be resistant to change (Chi & Roscoe, 2002) and relatively impervious to instruction (Konold, 1995). Using the SRA, researchers have found a gender gap, where women underperform in comparison to men, both by scoring lower on the correct reasoning scale and by demonstrating more misconceptions. In contrast, research using the CAOS has not examined gender as a factor of interest in statistical performance. However, it is becoming the gold standard in research on statistical competence in recent years (Lovett, Mayer, & Thille, 2008)

The inclusion of many different areas of understanding within those two research tools breaks from the tradition of published statistical reasoning research in psychology. Typically, those articles focused on a single aspect of statistical reasoning such as the law of large numbers (e.g., Fong, Krantz, and Nisbett, 1986), the need for comparison groups (e.g., Gray & Mill, 1990), or the importance of base rates in probability judgments (e.g., Bar-Hillel, 1980). Another important strength of these two scales is how they succeed at assessing students' understanding of statistical concepts without having recourse to calculations. Their multiple-choice format makes the SRA and the CAOS instruments of choice for classroom assessments and research.

As noted above, gender has emerged as a clear factor influencing statistical competence, as measured by the SRA. As found in mathematics, females tend to do more poorly than males. This finding has been called puzzling given that the gap occurs despite little or no difference in prior education (Tempelaar et al., 2006). It is thus not clear what causes this gender gap.

When research includes experience in statistics as the factor of interest, additional training does help augment the frequency and the quality of statistical answers. However, it is well known that even experts and professionals fail to achieve a perfect score on some statistical tasks (e.g., Hoffrage et al., 2002; Kahnemann et al., 1982). Thus, formal education does not appear sufficient on its own to ensure proper use of statistics outside the classroom. Yet, we know that "effective transfer is critical here because statistical reasoning is applicable across a wide variety of domains and in daily life; statistical reasoning skill is of little value if it can only be applied in the statistics classroom" (Lovett, 2001, p. 347).

An obvious gap in the literature is that when the impact of specific training and general class experience on statistical reasoning is examined, gender is typically not examined. Yet, we know that students of lower ability in statistics sometimes benefit more from training than those with higher ability (Quilici & Mayer, 1996). The interaction between gender and experience beyond that of a first course in statistics deserves attention given the particularly negative impact that the underperformance of women can have on their work prospects and success as citizens. Given this critical void in the literature, it was my goal in this dissertation to use a multipronged approach to shed light on the potential mechanisms underlying this gender difference in statistical reasoning, while at the same time examining the effect of experience. Besides experience, the three main factors examined included individual differences, stereotype threat, and task format.

Individual Differences

In Study 1, participants completed the SRA as well as multiple measures of thinking dispositions and cognitive ability to test the hypothesis that experience and individual differences would help explain the gender gap. As expected, the gender gap was observed on both types of reasoning scales (Correct reasoning and Misconceptions). Although added experience, as defined as the number of statistics courses taken in university, did lead to better performance on both the correct reasoning scale and on the misconceptions scale, it did not influence the size of the gender gap.

To further understand the contribution of individual differences, a confirmatory approach to structural equation modeling was used. Based on Stanovich's (2009) tri-partite model of reasoning, it was expected that both cognitive ability and thinking dispositions would influence statistical reasoning. Beyond the obvious role of cognitive ability, it was confirmed that thinking dispositions have a positive impact on statistical reasoning even when cognitive ability is controlled for. Those results provided evidence that Stanovich's (2009) tri-partite model of reasoning is appropriate and applicable to statistical reasoning, with the confirmation that the same process applies to both genders. As well, the addition of gender as a predictor in the model provided more information. Specifically, gender was found to significantly explain variation in statistical reasoning in three ways: Indirectly through its influence on cognitive ability; again indirectly but through its influence on thinking dispositions when cognitive ability is controlled for; and directly – thus above and beyond its indirect effect through cognitive ability and thinking dispositions. Hence, there are multiple ways in which performance in statistics can be influenced.

Given the significant role of thinking dispositions on statistical reasoning observed in Study 1, it was hypothesised that statistical performance may improve for women if one can influence the degree to which females are willing to cognitively engage with the task. This is in line with the idea that thinking dispositions, as opposed to cognitive ability, are relatively malleable and are thus a good target for interventions to improve reasoning performance (Baron, 1985; Stanovich, 2001). For studies 2, 3, and 4, efforts were focused on influencing thinking dispositions to increase engagement in the task, which, based on the results of the structural equation model, should improve performance.

Stereotype Threat

In an attempt to influence thinking dispositions, the phenomenon of stereotype threat was explored in Studies 2, 3 and 4. Stereotype threat is a situational phenomenon where the negative stereotype weighing on the reputation of a specific group reduces the ability of members of that targeted group to perform to their full potential. Reducing the threat is seen as a major way toward equal performance. Indeed, if the ability to focus on and engage in the task is restored, performance should increase for those affected by the threat. Thus, in Studies 2, 3, and 4, the self-affirmation technique promoted by Martens et al (2006) [in mathematics] and by Miyake et al (2010) [in physics] as an effective tool to counter the effect of stereotype threat was tested. Unlike some of the other methods suggested to counter stereotype threat (e.g., providing a female role model; providing a friendly environment), the great interest of this intervention lies in the potential it holds for use at time of testing by the stereotyped individual herself.

Unfortunately, value affirmation proved ineffective at influencing the performance and confidence of women on statistical reasoning tasks. This finding was replicated using two different tests (SRA, CAOS) in three separate studies, and held true even when the analysis was limited to those participants who agreed with a negative stereotype in statistics. This finding alone is inconsistent with the idea that level of agreement with a negative stereotype acts as a moderator to the effectiveness of the value affirmation exercise. Alternatively, this can be seen as evidence that stereotype threat is not a key mechanism underlying the observed gender differences in statistics. This possibility is consistent with a recent meta-analysis by Stoet and Geary (2012) that questions the importance of stereotype threat as a cause of a gender gap in mathematics. Some researchers also criticize stereotype threat as potentially being an effect limited to the laboratory (Sackett, Hardison, & Cullen, 2004; Sackett, Borneman, & Connellly, 2008), questioning the extent to which it applies to real-life settings. For instance, it is common for researchers to increase the saliency of gender prior to a test to create a performance gap and argue for the existence of a stereotype threat.

One potential limitation of the current studies is the fact that stereotype was not made salient through the manipulation of the context. However, this can be seen as both a limitation and strength. It is a limitation in the sense that the stereotype may need to be made more salient to truly impede women to perform at their best. On ecological validity grounds, it was chosen not to do this in the current series of experiments as it likely deviates significantly from real world experience. That is, one's gender typically is not made salient before one takes a statistics test.

Nonetheless, despite not manipulating the saliency of gender, the performance gap on the SRA (and CAOS in Study 4) was present. This could indicate one of two things. First, the stereotype in statistics may be largely prevalent, thus making priming the stereotype unnecessary in statistics to influence the performance of women – which would imply that value affirmation may not be a powerful enough intervention in statistics. Second, it could alternatively be that stereotype threat is not a factor of interest in statistical reasoning – which would explain why value affirmation did not have the expected effect.

However, not using a saliency manipulation helps answer the concern about the validity of the stereotype threat phenomenon outside the laboratory. In fact, the lack of impact of the value affirmation exercise in this context brings support to the idea that stereotype threat is not as important in real-life as what the literature of the past 20 years on the topic would like us to believe. It also concurs with the findings of Stoet and Geary (2012) that the beneficial effect of value affirmation is replicated in only 30 to 50% of the cases. Of course, their meta-analysis examined findings in mathematics, which may or may not apply directly to statistics.

It is also possible to criticize the format of the value affirmation exercise adopted in the current series of studies. For instance, Miyake et al. (2010) used a much longer set of questions in their field study. This higher level of engagement may be the key to an effective manipulation.

However, Martens et al. (2006) did report a positive effect of value affirmation with a shorter exercise, like that used in the current studies. In addition, controlling for level of engagement did not influence the pattern of results in the analyses. So, if the goal is to influence the disposition of women to engage in a statistical task, other interventions will have to be examined in future research. It addition, given the results obtained in the structural equation model, it is important to keep in mind that only about one-third of the effect in performance was related to the indirect effect of thinking dispositions. Thus, it should be expected that any intervention would have a limited impact on performance, and that any one intervention alone might not be sufficient.

Task Format and Mental Representations

As discussed above, both the SRA and the CAOS are designed to test the understanding of big statistical ideas without recourse to computations, formulas, or recall of definitions. The two tests rather focus on conceptual understanding. In contrast, statistics courses often follow the textbook, with chapters and notions tested sequentially and with little integration (Garfield, 1995). Students also tend to view what they learn as a set of isolated facts (Schoenfeld, 1987; Chi, 2005; diSessa, 2004). This is perhaps the reason females are often found to perform as well as males in the classroom. However, it does not specifically explain why a gender gap was found for the SRA and the CAOS.

Thus, for Studies 5 and 6, a different approach was used, this time focusing on the cognitive variable depicted in the structural equation model. Here, I explored the mental representations that participants held of statistical problems and tested the effect of using general versus constrained instructions on performance. These studies were designed to better understand the level of statistical sophistication of participants, and to examine whether males hold mental representations that are better assimilated than those of females, which would help explain their higher performance on the conceptual tasks used previously.

In these studies, a categorization task was used. Very few categorizing sorting tasks have been used to study statistical cognitive representations (Quilici & Mayer, 1996, 2002, Lavigne, Salkind & Yan, 2008, as well as Rabinowitz and Hogan, 2008, 2009, are the exceptions). In Studies 5 and 6, the seminal study of Chi et al. (1981) to study the development of mental representations was extended to the domain of statistics. In general, quantitative measures capturing the total number of sets created and the number of deviations from an ideal categorization did not inform the issue of changes in statistical sophistication through experience, or the issue of differences across genders. The qualitative measure, which was based on the quality of the rationales provided with each set, however, was much more informative. As with the previous studies, a change was observed only after two courses in statistics. That change was limited and, as noted by Shaughnessy (1992), very few people reached the pragmatic-statistical stage. However, there was an increasing tendency toward noticing the presence of statistical elements with added experience.

In Study 5, the task was *unconstrained:* there was no specification of how many sets participants should create, but only the instruction to classify the problems based on how they would solve them. Participants created anywhere between 2 and 12 sets. To eliminate the possibility that the number of sets created had influenced individuals' deviation scores and quality of rationales, and to examine the degree to which these unconstrained task instructions impacted the primary dependent variables, the study was replicated with a slight change to the task format. Simply, in Study 6, the task was *constrained*: participants were instructed to create exactly four sets, the number of sets dictated by the ideal solution. Once again, the total number of deviations from the ideal solution did not vary with experience and gender. However, this change in task format had a dramatic impact on the performance of females as reflected in the significant improvement of the quality of their rationales for their classifications. In contrast, males' rationales did not improve with the constrained task format.

Taken together, the findings from Study 5 and 6 are consistent with the idea that females seem to approach statistics word problems differently than males, which is consistent with the finding in mathematics. Specifically, despite being equally skilled at computing the solution when the problem is well defined for them, females have difficulty identifying and translating the strategy to use when faced with word problems (Quinn & Spencer, 2001). The current finding also suggests that the ability to identify the structural features of a problem is what offers the most room for growth in females. This finding implies that more effort must be deployed to have students in general (and females in particular) practice this skill. Unless students of statistics can identify the problem at hand, they are unlikely to conduct the proper calculations. Furthermore, they will learn to do well only what they practice doing (Garfield, 1995; Anderson, Corbett, & Conrad, 1989).

Implications and Future Directions

In summary, this dissertation aimed to shed light on the gender gap in statistics. Taken as a whole, Studies 1 to 4 established the presence of the same gender gap on two tests geared toward conceptual understanding in statistics, and value affirmation did not succeed as a manipulation to help close that gap. In contrast, even though it was not found to interact with gender to close the performance gap, on the SRA or the CAOS, incorporating experience as an additional factor was informative. With the inclusion of experience as an independent variable, it became obvious that a minimum of two courses is often necessary to create a significant change in performance. This finding is in line with a current argument among statistics educators: the need for a second course in statistics. Prominent statistics educators recently debated this topic at the first edition of the Electronic Conference on Teaching Statistics (eCOTS, 2012). Although the content of that second course is still disputed, high agreement exists in regard to the necessity of that second course (Isaacson & Schield, 2012). This need is obvious when considering that in all studies, even participants with two or more courses were far from achieving perfect performance. Indeed, the scores on the SRA and on the CAOS were far from ceiling (or floor when misconceptions are involved) in all four studies concerned. This is a typical finding with these tests, even though the low scores obtained by their students surprise even statistics instructors (delMas et al., 2007). delMas and his colleagues (2007) also found very little gains from pre-test to post-test on the CAOS. The same result was found in Study 2 using the SRA.

Studies 5 and 6 explored the influence of a cognitive factor – mental representations – in creating the gender gap. Critically, the format of the task was shown to be essential to the performance level of females, even leading them to perform better than males when the task was constrained. This finding holds great potential, both for future research and for in-class interventions. In research, it will be important to continue asking why females do not assimilate the big ideas in statistics as well as their male counterparts, even when they are able to obtain comparable grades on the typical in-class examinations. For instance, are they too focused on the details of the calculations to see the big picture? One avenue to explore this possibility could involve manipulating construal levels using Navon figures – where participants are instructed to report one of the two letters presented in a display characterized by a series of smaller (local) letters spatially arranged to form a larger (global) letter – prior to the completion of the CAOS or SRA. Perhaps, focusing on the global rather than on the local aspect of the task could help females perform better on those conceptual tests. For instance, past research with the local-global paradigm has demonstrated that using a local processing style makes it harder to pay attention to the relations between individual elements (Macrae & Lewis, 2002) and to judge complex stimuli (Dijkstra, van der Pligt, van Kleef, & Kerstholt, 2012). In addition, adopting a global processing style encourages integration of knowledge to make sense of a stimulus (Förster & Dannenberg, 2010).

At the practical level, it is important to ask how educators can design a female-friendly curriculum. To start with, educational research in statistics could make a point to report systematically the scores of males and females in their research. It is also important to establish the degree to which mathematics and statistics are similar and different, especially as the evolution of the computer tools that easily carry the mathematical calculations should now allow everyone to truly focus on the big ideas of statistics when encountering data.

This focus on the higher-level skills and knowledge of statistics has been recommended for the past two decades now that computer tools can easily compute the statistics of interest (Moore, 1992). Such change in focus requires a great commitment on the part of statistics educators, as they will have to think of ways to go beyond the compartmentalized textbook. Recent uses of online technology to complement in-class learning may offer part of the solution. For instance, Lovett, Mayer, and Thille (2008) followed the learning of students in three different versions of an introduction to statistics course: a traditional instructor-led course, a stand-alone web-based course, and a hybrid course. Those who accessed the lecture material online and met with the instructor in class to address the difficulties encountered through completion of the online activities, i.e., those in the hybrid format, progressed faster through the course material and obtained higher scores then those in traditional courses, both on the final exam and on the CAOS.

In the current context, more exercises – both online and in-class – could be developed to help students master the integration of topics studied across the multiple sessions, focusing specifically on the identification of the required solution strategy, which would first necessitate the identification of the underlying nature of the problem at hand. In keeping with the principles of learning in statistics, providing immediate feedback will help students consolidate their learning (Butler & Winne, 1995; Corbett & Anderson, 2001). Of course, the gender gap should be given more consideration when testing new pedagogical tools to ensure that all students will benefit equally from it.

One solution is unlikely to fit all. Efforts to improve attitudes of students, with the use of fun elements in class for instance (e.g., Lesser et al., in press), and efforts to ameliorate the quality of statistical education are all important. In a time when the field of statistics education is still defining itself, and when the world of data is growing exponentially, opportunities to contribute to the increased success of our citizens abound. Without a doubt, this challenge and opportunity to make a difference are truly exciting.

References

American Statistical Association (2005). *Guidelines for assessment and instruction in statistics education* (GAISE) *college report*. Alexandria, VA: ASA. www.amstat.org/education/gaise/

Anderson, J. R., Corbett, A. T., and Conrad, F. (1989). Skill acquisition and the LISP tutor. *Cognitive Science, 13*, 467-506.

Andersen, J.F., Norton, R.W., & Nussbaum, J.F. (1981). Three investigations exploring relationships between perceived teacher communication behaviors and student learning. *Communication Education, 30,* 377-392.

Andersen, J.F., & Withrow, J.G. (1981). The impact of lecturers‟ nonverbal expressiveness on improving mediated instruction. *Communication Education, 30,* 342-353.

Arbuckle J. L (2009). AMOS 18 user's guide. Chicago: AMOS Development Corporation.

Ashcraft, M. H., & Faust, M. W. (1994). Mathematics anxiety and mental arithmetic performance: An exploratory investigation. *Cognition and Emotion, 8*(2), 97-125. doi:10.1080/02699939408408931

Baron, J. (1985). *Rationality and intelligence*. New York: Cambridge University Press.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*(3), 211-233. doi:10.1016/0001-6918(80)90046-3

Ben-Zvi, & Garfield (2008). Introducing the Emerging Discipline of Statistics Education. *School Science and Mathematics, 108*(8), 355-361. doi: 10.1111/j.1949-8594.2008.tb17850.x

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Buckingham Shum, S. and Deakin Crick, R. (2012). Learning dispositions and transferable competencies: pedagogy, modelling and learning analytics. In: *2nd International*

Conference on Learning Analytics & Knowledge, 29 Apr - 02 May 2012, Vancouver, British Columbia, Canada (forthcoming).

Bullock, H.E., Harlow, L.L., & Mulaik, S.A. (1994). Causation issues in structural equation modeling research. *Structural Equation Modeling: A Multidisciplinary Journal, 1*(3). 253-267. doi:10.1080/10705519409539977

Byrnes, J. P., & Takahira, S. (1993). Explaining gender differences on SAT-math items. *Developmental Psychology, 29*(5), 805-810. doi:10.1037/0012-1649.29.5.805

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306-307.

Chi, M. T. H. (2005). Common Sense Conceptions of Emergent Processes: Why some misconceptions are robust. *Journal of the Learning Sciences, 14*, 161-199.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121-152.

Chi, M. & Roscoe, R. (2002). The processes and challenges of conceptual change. Limón, M. & Mason, L. (Eds.), *Reconsidering Conceptual Change: Issues in Theory and Practice*. (pp.3-27). doi:10.1007/0-306-47637-1_1

Cobb, G. (1992). Teaching statistics. *Heeding the Call for Change: Suggestions for Curricular Action,* (22), 3-43.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28-58.

delMas, R. C., & Liu, Y. (2007). Students' conceptual understanding of the standard deviation. In M. C. Lovett & P. Shah (Eds.), *Carnegie Mellon symposia on cognition. Thinking with data* (pp. 87-116). Mahwah, NJ: Lawrence Erlbaum Associates.

Dijkstra, K. A., van der Pligt, J., van Kleef, G. A., & Kerstholt, J. H. (2012). Deliberation versus intuition: Global versus local processing in judgment and choice. *Journal of Experimental Social Psychology, 48*(5), 1156-1161. doi:10.1016/j.jesp.2012.05.001

diSessa, A.A. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science*, 28, 843-900.

Dunbar, K. N., Fugelsang, J. A., & Stein, C. (2007). Do naïve theories ever go away? Using brain and behavior to understand changes in concepts. In M. C. Lovett & P. Shah (Eds.), *Carnegie Mellon symposia on cognition. Thinking with data* (pp. 193-205). Mahwah, NJ: Lawrence Erlbaum Associates.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice.* (pp. 9-26). New York: W. H. Reeman and Company.

Ennis, R. H. (1996). Critical thinking dispositions: Their nature and assessability. *Informal Logic. 18*(2&3), 129-147.

Fischbein, E. (1987). *Mathematics education library. Intuition in science and mathematics: An educational approach.* Dordrecht, Netherlands: D Reidel.

Förster, J., & Dannenberg, L. (2010). GLOMOsys: A systems account of global versus local processing. *Psychological Inquiry, 21*(3), 175-197. doi:10.1080/1047840X.2010.487849

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25-42.

Garfield, J. (1991). Evaluating Students' Understanding of Statistics: Development of the Statistical Reasoning Assessment. In *Proceedings of the Thirteenth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Volume 2.* Blacksburg, VA, pp. 1-7.

Garfield, J. (1995). How students learn statistics. *International Statistical Review, 63*, 25–34.

Garfield, J. (1998) The Statistical Reasoning Assessment: Development and Validation of a Research Tool. In *L Pereira-Mendoza (Ed.) Proceedings of the Fifth International Conference on Teaching Statistics*, pp. 781-786. Voorburg, The Netherlands: International Statistical Institute.

Garfield, J. (2003). Assessing Statistical Reasoning. *Statistics Education Research Journal [Online], 2(1),* 22-38.

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

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning, 2*(1-2), 99-125. doi:10.1207/S15327833MTL0202_5

Garfield, J., & Gal, I. (1999). Assessment and Statistics Education: Current Challenges and Directions, *International Statistical Review / Revue Internationale de Statistique*, *67*(1), 1-12. Stable URL: http://www.jstor.org/stable/1403562

Goldman, R.N., & Weinberg, J.S. (1985). Statistics: an introduction. Englewood Cliffs, N.J.: Prentice-Hall.

Gorham, J. (1988). The relationship between verbal teacher immediacy behaviors and student learning. *Communication Education, 37,* 40-53.

Gray, T., & Mill, D. (1990). Critical abilities, graduate education (Biology vs. English), and belief in unsubstantiated phenomena. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 22*(2), 162-172. doi:10.1037/h0078899

Hawkins, A. (1997). Myth-conceptions. In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. vii-viii). Voorburg, The Netherlands: International Statistical Institute.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science, 290*(5500), 2261-2262. doi:10.1126/science.290.5500.2261

Hogan, T., & Rabinowitz, M. (2009). Teacher expertise and the development of a problem representation. *Educational Psychology, 29*(2), 153-169. doi:10.1080/01443410802613301

Howell, D.C. (2010) *Statistical methods for psychology*. Belmont, CA: Cengage Wadsworth.

Isaacson, M., & Schield, M. (2012, May 15). *A "Second" Statistics Course is Needed: What should it be?* Paper presented at the Electronic Conference On Teaching Statistics (eCOTS). http://www.causeweb.org/ecots/breakouts/7/

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science, 16*(3), 175-179. doi:10.1111/j.0956-7976.2005.00799.x

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26. http://www.jstor.org/stable/3700093

Joreskog, K. G., & Sorbom, D. (1989). *LISREL7 users' reference guide*. Chicago, IL: Scientific Software International.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.

Kline, R. B. (2011). Principles and practice of structural equation modeling (3rd ed.). New York: Guilford.

Koenig, A.M., & Eagly, A.H. (2005). Stereotype Threat in Men on a Test of Social Sensitivity. *Sex Roles, 52*(7–8): 489–496. doi:10.1007/s11199-005-3714-x

Lakens, D., Schneider, I. K., Jostmann, N. B., & Schubert, T. W. (2011). Telling things apart: The distance between response keys influences categorization times. *Psychological Science, 22*(7), 887-890. doi:10.1177/0956797611412391

Lesser, L.M., Wall, A., Carver, R., Pearl, D.K., Martin, N., Kuiper, S., Posner, M.A., Erickson, P., Liao, S.-M., Albert, J., & Weber, J.J. (in press). Using Fun in the Statistics Classroom: An Exploratory Study of College Instructors' Hesitations and Motivations. *Journal of Statistics Education*.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123-1135. doi:10.1037/a0021276

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21,* 37–44.

Lovett, M. C. (2001). A Collaborative convergence on studying reasoning processes: A case study in statistics. In S. Carver, & D. Klahr (Eds.) *Cognition and instruction: Twenty-five years of progress* (pp. 347-384). Mahwah, NJ: Erlbaum.

Lovett, M., Meyer, O. and Thille, C. (2008). The Open Learning Initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal of Interactive Media in Education*, *14*. http://jime.open.ac.uk/2008/14

Macrae, C. N., & Lewis, H. L. (2002). Do I know you?: Processing orientation and face recognition. *Psychological Science, 13*(2), 194-196. doi:10.1111/1467-9280.00436

Martens, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology, 42*(2), 236-243.

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L ., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science, 330*(6008), 1234-1237. doi:10.1126/science.1195996

Mulhern, G., & Wylie, J. (2006). Mathematical prerequisites for learning statistics in psychology: Assessing core skills of numeracy and mathematical reasoning among undergradutes. *Psychology Learning & Teaching, 5*(2), 119-132. doi:10.2304/plat.2005.5.2.119

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Parsons, S., & Bynner, J. (1997). Numeracy and employment. *Education and Training*, 39, 43-51.

Polya, G. (1945). *How to solve it; a new aspect of mathematical method*. Princeton, NJ: Princeton University Press.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88*(1), 144-161. doi:10.1037/0022-0663.88.1.144

Quilici, J. L., & Mayer, R. E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Applied Cognitive Psychology, 16*(3), 325-342. doi:10.1002/acp.796

Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues, 57*(1), 55-71. doi:10.1111/0022-4537.00201

Rabinowitz, M., & Hogan, T. M. (2008). Experience and problem representation in statistics. *The American Journal of Psychology, 121*(3), 395-407. doi:10.2307/20445474

Reyna, V. F., Nelson, W., Han, P., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*, 943-973. doi:10.1037/a0017327.

Ritchhart, R. (2001). From IQ to IC: A dispositional view of intelligence. *Roeper Review: A Journal on Gifted Education, 23*(3), 143-150. doi:10.1080/02783190109554086

Sá, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 91*(3), 497-510.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*(4), 215-227. doi:10.1037/0003-066X.63.4.215

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On Interpreting Stereotype Threat as Accounting for African American-White Differences on Cognitive Tests. *American Psychologist, 59*(1), 7-13. doi:10.1037/0003-066X.59.1.7

Schield, M. (2005). *Five Percentage Table Survey*. W. M. Keck Statistical Literacy Project. http://www.statlit.org/gc/p3/PrcntgTblSurvey.aspx

Schmader, T., Johns, M., & Barquissau, M. (2004). The Costs of Accepting Gender Differences: The Role of Stereotype Endorsement in Women's Experience in the Math Domain. *Sex Roles, 50*(11-12), 835-850. doi:10.1023/B:SERS.0000029101.74557.a0

Schoenfeld, A. H. (1987). What's all the fuss about meta-cognition? In A. H. Schoenfeld (Ed.), *Cognitive Science and Mathematics Education*. Hillsdale, NJ: Erlbaum.

Shapiro, J. R., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles, 66*(3-4), 175-183. doi:10.1007/s11199-011-0051-0

Shaughnessy, M. J. (1992). Research in probability and statistics: Reflections and directions. In D. Grouws (Ed.), *Handbook on research in mathematics education* (pp. 465-494). New York: Macmillan.

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition, 34*(3), 619-632.

Stanovich, K.E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In Evans, J.S.B.T. & Frankish, K. (Eds.), *In two minds: Dual processes and beyond*. (pp. 55-88). New York, NY, US: Oxford University Press.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*(2), 342-357. doi:10.1037/0022-0663.89.2.342

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*(2), 161-188.

Stanovich, K. E., & West, R. F. (1998b). Individual differences in framing and conjunction efforts. *Thinking & Reasoning, 4*(4), 289-317. doi:10.1080/135467898394094

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning, 13*(3), 225-247. doi:10.1080/13546780600780796

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of african americans. *Journal of Personality and Social Psychology, 69*(5), 797-811. doi:10.1037/0022-3514.69.5.797

Stone, J., Perry, Z. W., & Darley, J. M. (1997). "White men can't jump": Evidence for the perceptual confirmation of racial stereotypes following a basketball game. *Basic and Applied Social Psychology, 19*(3), 291-306. doi:10.1207/15324839751036977

Taylor, V. J., & Walton, G. M. (2011). Stereotype threat undermines academic learning. *Personality and Social Psychology Bulletin, 37*(8), 1055-1067. doi:10.1177/0146167211406506

Tempelaar, D.T., Gijselaers, W.H. & Schim van der Loeff, S. (2006). Puzzles in Statistical Reasoning. *Journal of Statistics Education*, 14 (1). www.amstat.org/publications/jse/v14n1/tempelaar.html

Tishman, S., & Andrade. A. (1995). Thinking dispositions: A review of current theories, practices, and issues. (Internal report). Cambridge, MA: Project Zero: Harvard University Graduate School of Education.

Toplak, M. E., & Stanovich, K. E. (2003). Associations between myside bias on an informal reasoning task and amount of post-secondary education. *Applied Cognitive Psychology, 17*(7), 851-860. doi:10.1002/acp.915

Toplak, M. E., & Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 94*(1), 197-209. doi:10.1037/0022-0663.94.1.197

Viswanathan, M. (1993). Measurement of individual differences in preference for numerical information. *Journal of Applied Psychology, 78*(5), 741-752. doi:10.1037/0021-9010.78.5.741

Wallman, K.K. (1993). Enhancing statistical literacy: enriching our society. *Journal of the American Statistical Association*, *88*(421), 1-8.

Williams, A. S. (2010). Statistics anxiety and instructor immediacy. *Journal of Statistics Education, 18*(2), 1-18. http://www.amstat.org/publications/jse/v18n2/williams.pdf

Wonderlic Inc. (1999). *Wonderlic's personnel test manual and scoring guide.* Libertyville, IL: Wonderlic.

Woody, E. (2011). An SEM Perspective on Evaluating Mediation: What Every Clinical Researcher Needs to Know. *Journal of Experimental Psychopathology, 2*(2), 210-251. doi:10.5127/jep.010410

APPENDICES

Appendix A – Statistical Reasoning Assessment (including subscales)

A1. List of Correct Reasoning and Misconceptions subscales on the SRA (from Tempelaar et al, 2006)

Correct Reasoning Subscales:

Subscale #	Outcome assessed
CR1:	*Correctly interprets probabilities*. Assesses the understanding and use of ideas of randomness, chance to make judgments about uncertain events
CR2:	*Understands how to select an appropriate average*. Assesses the understanding what measures of center tell about a data set, and which are best to use under different conditions.
CR3:	*Correctly computes probability, both understanding probabilities as ratios, and using combinatorial reasoning*. Assesses the knowledge that in uncertain events not all outcomes are equally likely, and how to determine the likelihood of different events using an appropriate method.
CR4:	*Understands independence*.
CR5:	*Understands sampling variability*
CR6:	*Distinguishes between correlation and causation*. Assesses the knowledge that a strong correlation between two variables does not mean that one causes the other.
CR7:	*Correctly interprets two-way tables*. Assesses the knowledge how to judge and interpret a relationship between two variables, knowing how to examine and interpret a two-way table.
CR8:	*Understands the importance of large samples*. Assesses the knowledge of how samples are related to a population and what may be inferred from a sample; knowing that a larger, well chosen sample will more accurately represent a population; being cautious when making inferences made on small samples.

Misconception scales:

Subscale #	Outcome assessed
MISC1	*Misconceptions involving averages*. This category includes the following pitfalls: averages are the most common number; failing to take outliers into consideration when computing the mean; comparing groups on their averages only; and confusing mean with median.
MISC2	*Outcome orientation*. Students use an intuitive model of probability that lead them to make yes or no decisions about single events rather than looking at the series of events; see Konold (1989).
MISC3	*Good samples have to represent a high percentage of the population*. Size of the sample and how it is chosen is not important, but it must represent a large part of the population to be a good sample.
MISC4	*Law of small numbers*. Small samples best resemble the populations from which they are sampled, so are to be preferred over larger samples.
MISC5	*Representativeness misconception*. In this misconception the likelihood of a sample is estimated on the basis how closely it resembles the population. Documented in Kahneman,

	Slovic, & Tversky (1982).
MISC6	*Correlation implies causation*.
MISC7	*Equiprobability bias*. Events of unequal chance tend to be viewed as equally likely; see Lecoutre (1992).
MISC8	*Groups can only be compared if they have the same size*.

A2. Statistical Reasoning Assessment (SRA)

Question 1. [CR2/CR7/MISC1][5] A small object was weighed on the same scale, separately by nine students, in a science class. The weights (in grams) recorded by each student are shown below.

6.2 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.2

The students want to determine as accurately as they can the actual weight of this object. Of the following methods, which would you recommend they use?

a. Use the most common number, which is 6.2.
b. Use the 6.15 since it is the most accurate weighing.
c. Add up the 9 numbers and divide by 9.
d. Throw out the 15.3, add up the other 8 numbers and divide by 8.

Indicate how confident you are that you correctly answered the previous question.[6]

```
  1----------2----------3----------4----------5----------6
Very                                                    Very
Low                                                     High
```

Question 2. [CR1/MISC2] The following message is printed on a bottle of prescription medication:
WARNING: For applications to skin areas there is a 15% chance of developing a rash. If a rash develops, consult your physician.

Which of the following is the best interpretation of this warning?

a. Don’t use the medication on your skin, there’s a good chance of developing a rash.
b. For application to the skin, apply only 15% of the recommended dose.
c. If a rash develops, it will probably involve only 15% of the skin.
d. About 15 of 100 people who use this medication develop a rash.
e. There is hardly a chance of getting a rash using this medication.

Question 3. [CR1/ MISC2] The Springfield Meteorological Center wanted to determine the accuracy of their weather forecasts. They searched their records for those days when the forecaster had reported a 70% chance of rain. They compared these forecasts to records of whether or not it actually rained on those particular days.

The forecast of 70% chance of rain can be considered very accurate if it rained on:

a. 95% - 100% of those days.
b. 85% - 94% of those days.
c. 75% - 84% of those days.
d. 65% - 74% of those days.
e. 55% - 64% of those days.

Question 4. [CR2] A teacher wants to change the seating arrangement in her class in the hope that it will increase the number of comments her students make. She first decides to see how many comments students make with the current seating arrangement. A record of the number of comments made by her 8 students during one class period is

[5] These codes refer to the notions tested in the question and correspond to the subscales presented on pp.124-125.
[6] This question was inserted after each of the 20 questions on the SRA to prompt participants to report their level of confidence. For simplicity, the question will not be repeated in this appendix.

shown below.

Student Initials								
	A.A.	R.F.	A.G.	J.G.	C.K.	N.K.	J.L.	A.W.
Number of comments	0	5	2	22	3	2	1	2

She wants to summarize this data by computing a typical number of comments made that day. Of the following methods, which would you recommend she use?

a. Use the most common number, which is 2.
b. Add up the 8 numbers and divide by 8.
c. Throw out the 22, add up the other 7 numbers and divide by 7.
d. Throw out the 0, add up the other 7 numbers and divide by 7.

Question 5. [CR7] A new medication is being tested to determine its effectiveness in the treatment of eczema, an inflammatory condition of the skin. Thirty patients with eczema were selected to participate in the study. The patients were randomly divided into two groups. Twenty patients in an experimental group received the medication, while ten patients in a control group received no medication. The results after two months are shown below.

	Experimental group (Medication)	Control group (No Medication)
Improved	8	2
No Improvement	12	8

Based on the data, I think the medication was:

1. Somewhat effective

2. Basically ineffective

<u>If you chose option 1,</u> select the one explanation below that best describes your reasoning.

a. 40% of the people (8/20) in the experimental group improved.

b. 8 people improved in the experimental group while only 2 improved in the control group.

c. In the experimental group, the number of people who improved is only 4 less than the number who didn't improve (12-8), while in the control group, the difference is 6 (8-2).

d. 40% of the patients in the experimental group improved (8/20), while only 20% improved in the control group.

<u>If you chose option 2,</u> select the one explanation below that best describes your reasoning.

a. In the control group, 2 people improved even without medication.

b. In the experimental group, more people didn't get better than did (12 vs 8).

c. The difference between the numbers who improved and didn't improve is about the same in each group (4 vs 6).

d. In the experimental group, only 40% of the patients improved (8/20).

Question 6. [CR8/ MISC8] Listed below are several possible reasons one might question the results of the experiment described above. Circle on your answer sheet every statement that you agree with.

a. It’s not legitimate to compare the two groups because there are different numbers of patients in each group.
b. The sample of 30 is too small to permit drawing conclusions.
c. The patients should not have been randomly put into groups, because the most severe cases may have just by chance ended up in one of the groups
d. I’m not given enough information about how doctors decided whether or not patients improved. Doctors may have been biased in their judgments.
e. I don’t agree with any of these statements

Question 7. [MISC3] A marketing research company was asked to determine how much money teenagers (ages 13-19) spend on recorded music (cassette tapes, CDs and records). The company randomly selected 80 malls located around the country. A field researcher stood in a central location in the mall and asked passers-by who appeared to be the appropriate age to fill out a questionnaire. A total of 2, 050 questionnaires were completed by teenagers. On the basis of this survey, the research company reported that the average teenager in this country spends $155 each year on recorded music.

Listed below are several statements concerning this survey. Circle on your answer sheet every statement that you agree with.

a. The average is based on teenagers’ estimates of what they spend and therefore could be quite different from what teenagers actually spend
b. They should have done the survey at more than 80 malls if they wanted an average based on teenagers throughout the country.
c. The sample of 2, 050 teenagers is too small to permit drawing conclusions about the entire country.
d. They should have asked teenagers coming out of music stores.
e. The average could be a poor estimate of the spending of all teenagers given that teenagers were not randomly chosen to fill out the questionnaire.
f. The average could be a poor estimate of the spending of all teenagers given that only teenagers in malls were sampled
g. Calculating an average in this case is inappropriate since there is a lot of variation in how much teenagers spend.
h. I don’t agree with any of these statements.

Question 8. [CR3] Two containers, labelled A and B, are filled with red and blue marbles in the following quantities:

Container	Red	Blue
A	6	4
B	60	40

Each container is shaken vigorously. After choosing one of the containers, you will reach in and, without looking, draw out a marble. If the marble is blue, you win $50. Which container gives you the best chance of drawing a blue marble?

a. Container A (with 6 red and 4 blue)
b. Container B (with 60 red and 40 blue)
c. Equal chances from each container

Question 9. [CR4/ MISC5] Which of the following sequences is most likely to results from flipping a fair coin 5 times?

a. H H H T T

b. T H H T H
c. T H T T T
d. H T H T H
e. All four sequences are equally likely.

Question 10. [CR4/ MISC5] Select one or more explanations for the answer you gave for the item above.

a. Since the coin is fair, you ought to get roughly equal numbers of heads and tails.
b. Since coin flipping is random, the coin ought to alternate frequently between landing heads and tails.
c. Any of the sequences could occur.
d. If you repeatedly flipped a coin five times, each of these sequences would occur about as often as any other sequence.
e. If you get a couple of heads in a row, the probability of a tails on the next flip increases.
f. Every sequence of five flips has exactly the same probability of occurring.

Question 11. [CR4/MISC2/MISC5] Which of the following sequences is <u>least</u> likely to result from flipping a fair coin 5 times?

a. H H H T T
b. T H H T H
c. T H T T T
d. H T H T H
e. All four sequences are equally unlikely

Question 12. [CR8/MISC2/MISC4] The Caldwells want to buy a new car, and they have narrowed their choices to a Buick or a Oldsmobile. They first consulted an issue of <u>Consumer Reports</u>, which compared rates of repair for various cars. Records of repairs done on 400 cars of each type showed somewhat fewer mechanical problems with the Buick than with the Oldsmobile.
The Caldwells then talked to three friends, two Oldsmobile owners, and one former Buick owner. Both Oldsmobile owners reported having a few mechanical problems, but nothing major. The Buick owner, however, exploded when asked how he liked his car:
First, the fuel injection went out - $250 bucks. Next, I started having trouble with the rear end and had to replace it. I finally decided to sell it after the transmission went. I'd never buy another Buick.

The Caldwells want to buy the car that is less likely to require major repair work. Given what they currently know, which car would you recommend that they buy?

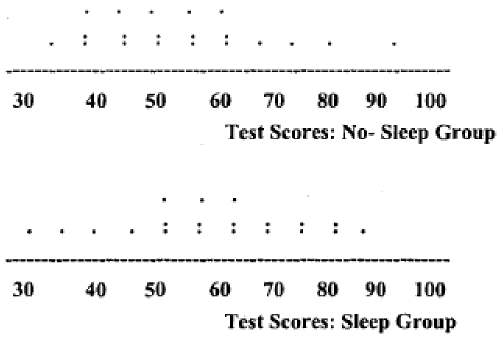
a. I would recommend that they buy the Oldsmobile, primarily because of all the trouble their friend had with his Buick. Since they haven't heard similar horror stories about the Oldsmobile, they should go with it.
b. I would recommend that they buy the Buick in spite of their friend's bad experience. That is just one case, while the information reported in <u>Consumer Reports</u> is based on many cases. And according to that data, the Buick is somewhat less likely to require repairs.
c. I would tell them that it didn't matter which car they bought. Even though one of the models might be more likely than the other to require repairs, they could still, just by chance, get stuck with a particular car that would need a lot of repairs. They may as well toss a coin to decide.

Question 13. [CR3/ MISC2/MISC7] Five faces of a fair die are painted black, and one face is painted white. The die is rolled six times. Which of the following results is more likely?

a. Black side up on five of the rolls; white side up on the other roll
b. Black side up on all six rolls
c. <u>a</u> and <u>b</u> are equally likely

Question 14. [CR5/ MISC4] Half of all newborns are girls and half are boys. Hospital A records and average of 50 births a day. Hospital B records an average of 10 births a day. On a particular day, which hospital is more likely to record 80% or more female births?

a. Hospital A (with 50 births a day)
b. Hospital B (with 10 births a day)
c. The two hospitals are equally likely to record such an event



Question 15. [CR5/MISC1] Forty college students participated in a study of the effect of sleep on test scores. Twenty of the students volunteered to stay up all night studying the night before the test (no-sleep group). The other 20 students (the control group) went to bed by 11:00pm on the evening before the test. The test scores for each group are shown in the graphs below. Each dot on the graph represents a particular student's score. For example, the two dots above the 80 in the bottom graph indicate that two students in the sleep group scored 80 on the test.

Examine the two graphs carefully. Then choose from the 6 possible conclusions listed below the one you <u>most</u> agree with.

a. The no-sleep group did better because none of these students scored below 40 and the highest score was achieved by a student in this group.
b. The no-sleep group did better because its average appears to be a little higher than the average of the sleep group.
c. There is no difference between the two groups because there is considerable overlap in the scores of the two groups.
d. There is no difference between the two groups because the difference between their averages is small compared to the amount of variation in the scores.
e. The sleep group did better because more students in this group scored 80 or above.
f. The sleep group did better because its average appears to be a little higher than the average of the no-sleep group.

Question 16. [CR6/MISC2/MISC6] For one month, 500 elementary students kept a daily record of the hours they spent watching television. The average number of hours per week spent watching television was 28. The researchers conducting the study also obtained report cards for each of the students. They found that the students who did well in school spent less time watching television than those students who did poorly. Listed below are several possible statements concerning the results of this research. Circle on your answer sheet every statement that you agree with.

a. The sample of 500 is too small to permit drawing conclusions.
b. If a student decreased the amount of time spent watching television, his or her performance in school would improve.
c. Even though students who did well watched less television, this doesn't necessarily mean that watching television hurts school performance.

d. One month is not a long enough period of time to estimate how many hours the students really spend watching television.
e. The research demonstrates that watching television causes poorer performance in school
f. I don't agree with any of these statements

Question 17. [CR2/ MISC1] The school committee of a small town wanted to determine the average number of children per household in their town. They divided the total number of children in the town by 50, the total number of households. Which of the following statements must be true if the average children per household is 2.2?

a. Half the households in the town have more than 2 children.
b. More households in the town have 3 children than have 2 children.
c. There are a total of 110 children in the town.
d. There are 2.2 children in the town for every adult.
e. The most common number of children in a household is 2.
f. None of the above.

Question 18. [CR3/ MISC7] When two dice are simultaneously thrown it is possible that one of the following two results occurs: *Result 1*: A 5 and a 6 are obtained. *Result 2*: A 5 is obtained twice.

Select the response that you agree with the most:

a. The chance of obtaining each of these results is equal.
b. There is more chance of obtaining result 1.
c. There is more chance of obtaining result 2.
d. It is impossible to give an answer.

Question 19. [CR3/MISC7] When three dice are simultaneously thrown, which of the following results is MOST LIKELY to be obtained?

a. *Result 1*: "A 5, a 3 and a 6"
b. *Result 2*: "A 5 three times"
c. *Result 3*: "A 5 twice and a 3"
d. All three results are equally likely

Question 20. [CR3/ MISC7] When three dice are simultaneously thrown, which of these three results is LEAST LIKELY to be obtained?

a. *Result 1*: "A 5, a 3 and a 6"
b. *Result 2*: "A 5 three times"
c. *Result 3*: "A 5 twice and a 3"
d. All three results are equally unlikely.

Appendix B – Measures of Individual Differences

B1. Thinking Dispositions: Preference for Numerical Information (PNI) Scale & Level of Endorsement of the Stereotype

1. I enjoy work that requires the use of numbers. ________
2. I think quantitative information is difficult to understand. ________
3. I find it satisfying to solve day-to-day problems involving numbers. ________
4. Numerical information is very useful in everyday life. ________
5. I prefer not to pay attention to information involving numbers. ________
6. I think more information should be available in numerical form. ________
7. I don't like to think about issues involving numbers. ________
8. Numbers are not necessary for most situations. ________
9. Thinking is enjoyable when it does not involve quantitative information. ________
10. I like to make calculations using numerical information. ________
11. Quantitative information is vital for accurate decisions. ________
12. I enjoy thinking about issues that do not involve numerical information. ________
13. Understanding numbers is as important in daily life as reading or writing. ________
14. I easily lose interest in graphs, percentages, and other quantitative information. ________
15. I don't find numerical information to be relevant for most situations. ________
16. I think it is important to learn and use numerical information to make well-informed decisions. __
17. Numbers are redundant for most situations. ________
18. It is a waste of time to learn information containing a lot of numbers. ________
19. I like to go over numbers in my mind. _______
20. It helps me to think if I put down information as numbers. ________

*21. According to my own personal beliefs, I expect men to generally do better in math than women.
*22. According to my own personal beliefs, I expect men to generally do better in statistics than women.

**These questions were included to probe levels of agreement with the stereotypes. They were used in studies 2,3, and 4, but not in Study 1. In Phase 1 of Study 2, these two questions were included at the end of the SRA instead of at the end of the PNI.*

B2. Thinking Dispositions: Need for Cognition (NC) Scale

1. I would prefer complex problems to simple ones. _______
2. I like to have the responsibility of handling a situation that requires a lot of thinking. _______
3. Thinking is not my idea of fun. _______
4. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities. _______
5. I try to anticipate and avoid situations where there is a likely chance that I will have to think in depth about something. _______
6. I find satisfaction in deliberating hard and for long hours. _______
7. I only think as hard as I have to. _______
8. I prefer to think about small, daily projects to long-term ones. _______
9. I like tasks that require little thought once I've learned them. _______
10. The idea of relying on thought to make my way to the top appeals to me. _______
11. I really enjoy a task that involves coming up with new solutions to problems. _______
12. Learning new ways to think doesn't appeal to me very much. _______
13. I prefer my life to be filled with puzzles that I must solve. _______
14. The notion of thinking abstractly is appealing to me. _______
15. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought. ______
16. I feel relief rather than satisfaction after completing a task that required a lot of mental effort. ____
17. It's enough for me that something gets the job done; I don't care how or why it works. _______
18. I usually end up deliberating about issues even when they do not affect me personally. _______

B3. Thinking Dispositions: Actively Open-Minded Thinking (AOT) Scale

1. Even though freedom of speech for all groups is a worthwhile goal, it is unfortunately necessary to restrict the freedom of certain political groups. (Reflected)
2. What beliefs you hold have more to do with your own personal character than the experiences that may have given rise to them. (Reflected)
3. I tend to classify people as either for me or against me. (Reflected)
4. A person should always consider new possibilities.
5. There are two kinds of people in this world: those who are for the truth and those who are against the truth. (Reflected)
6. Changing your mind is a sign of weakness. (Reflected)
7. I believe we should look to our religious authorities for decisions on moral issues. (Reflected)
8. I think there are many wrong ways, but only one right way, to almost anything. (Reflected)
9. It makes me happy and proud when someone famous holds the same beliefs that I do. (Reflected)
10. Difficulties can usually be overcome by thinking about the problem, rather than through waiting for good fortune.
11. There are a number of people I have come to hate because of the things they stand for. (Reflected)
12. Abandoning a previous belief is a sign of strong character.
13. No one can talk me out of something I know is right. (Reflected)
14. Basically, I know everything I need to know about the important things in life. (Reflected)
15. It is important to persevere in your beliefs even when evidence is brought to bear against them. (Reflected)
16. Considering too many different opinions often leads to bad decisions. (Reflected)
17. There are basically two kinds of people in this world, good and bad. (Reflected)
18. I consider myself broad-minded and tolerant of other people's lifestyles.
19. Certain beliefs are just too important to abandon no matter how good a case can be made against them. (Reflected)
20. Most people just don't know what's good for them. (Reflected)
21. It is a noble thing when someone holds the same beliefs as their parents. (Reflected)
22. Coming to decisions quickly is a sign of wisdom. (Reflected)
23. I believe that loyalty to one's ideals and principles is more important than "open-mindedness." (Reflected)
24. Of all the different philosophies which exist in the world there is probably only one which is correct. (Reflected)
25. My beliefs would not have been very different if I had been raised by a different set of parents. (Reflected)
26. If I think longer about a problem I will be more likely to solve it.
27. I believe that the different ideas of right and wrong that people in other societies have may be valid for them.
28. Even if my environment (family, neighborhood, schools) had been different, I probably would have the same religious views. (Reflected)
29. There is nothing wrong with being undecided about many issues.
30. I believe that laws and social policies should change to reflect the needs of a changing world.
31. My blood boils over whenever a person stubbornly refuses to admit he's wrong. (Reflected)
32. I believe that the "new morality" of permissiveness is no morality at all. (Reflected)
33. One should disregard evidence that conflicts with your established beliefs. (Reflected)
34. Someone who attacks my beliefs is not insulting me personally.
35. A group which tolerates too much difference of opinion among its members cannot exist for long. (Reflected)
36. Often, when people criticize me, they don't have their facts straight. (Reflected)
37. Beliefs should always be revised in response to new information or evidence.
38. I think that if people don't know what they believe in by the time they're 25, there's something wrong with them. (Reflected)
39. I believe letting students hear controversial speakers can only confuse and mislead them. (Reflected)
40. Intuition is the best guide in making decisions. (Reflected)
41. People should always take into consideration evidence that goes against their beliefs.

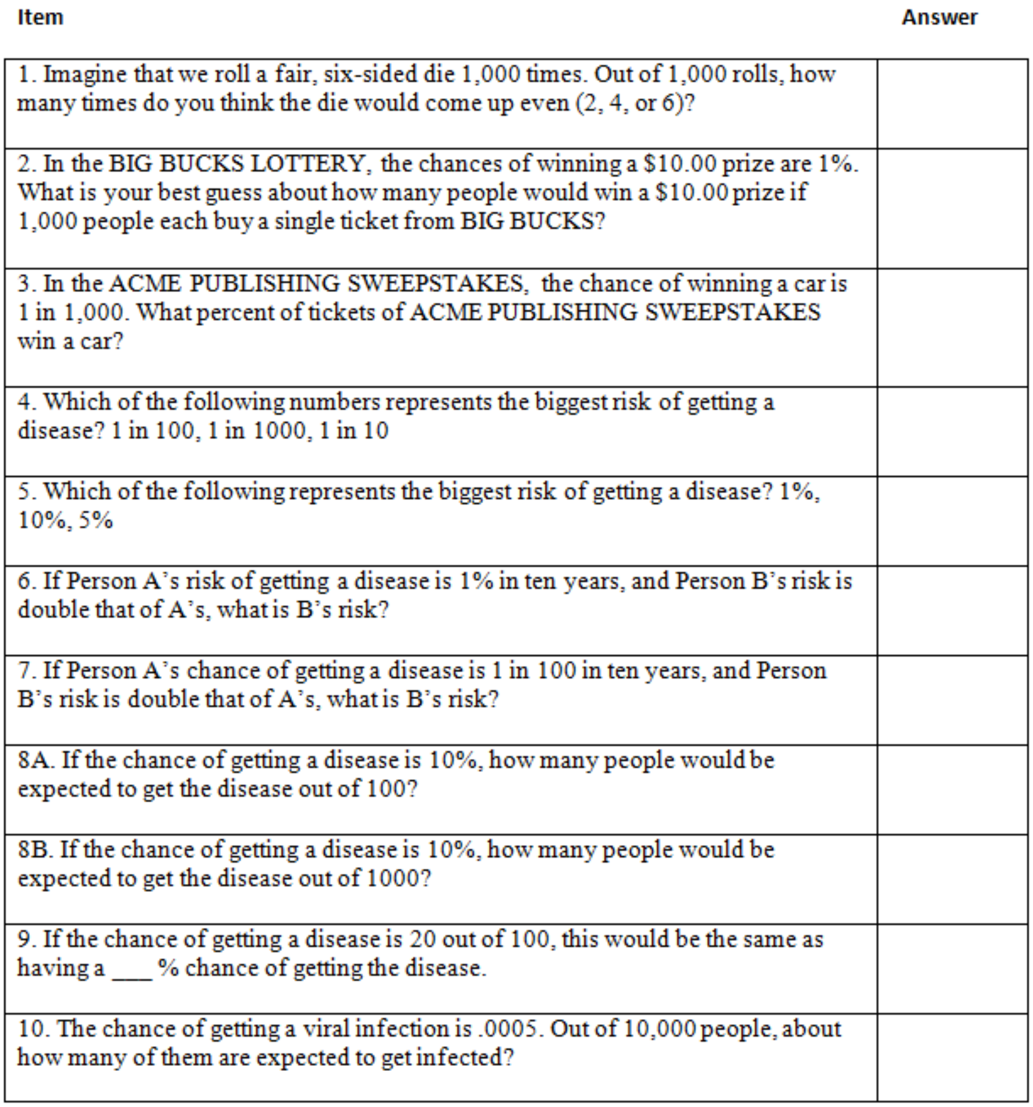
B4. Cognitive Ability: Cognitive Reflection Test (CRT)

1. A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? ____ cents

2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? ____ minutes

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? ____ days

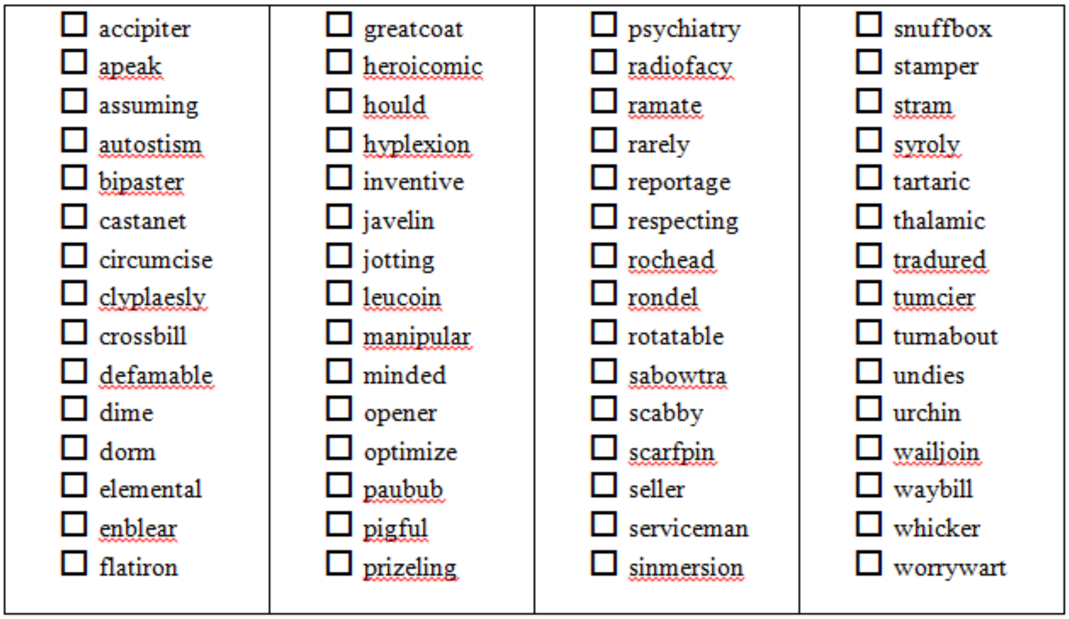
B5. Cognitive Ability: Numeracy (NUM) Scale



Item	Answer
1. Imagine that we roll a fair, six-sided die 1,000 times. Out of 1,000 rolls, how many times do you think the die would come up even (2, 4, or 6)?	
2. In the BIG BUCKS LOTTERY, the chances of winning a $10.00 prize are 1%. What is your best guess about how many people would win a $10.00 prize if 1,000 people each buy a single ticket from BIG BUCKS?	
3. In the ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1,000. What percent of tickets of ACME PUBLISHING SWEEPSTAKES win a car?	
4. Which of the following numbers represents the biggest risk of getting a disease? 1 in 100, 1 in 1000, 1 in 10	
5. Which of the following represents the biggest risk of getting a disease? 1%, 10%, 5%	
6. If Person A's risk of getting a disease is 1% in ten years, and Person B's risk is double that of A's, what is B's risk?	
7. If Person A's chance of getting a disease is 1 in 100 in ten years, and Person B's risk is double that of A's, what is B's risk?	
8A. If the chance of getting a disease is 10%, how many people would be expected to get the disease out of 100?	
8B. If the chance of getting a disease is 10%, how many people would be expected to get the disease out of 1000?	
9. If the chance of getting a disease is 20 out of 100, this would be the same as having a ___ % chance of getting the disease.	
10. The chance of getting a viral infection is .0005. Out of 10,000 people, about how many of them are expected to get infected?	

B6. Cognitive Ability: Vocabulary Checklist-with-Foils (VOC)

For this task, you will see 60 letter strings. Some are actual words and others are not actual words. Read through the list. Every time you know that it is a word, put a checkmark in the box beside it. Do not use a dictionary.



□ accipiter
□ apeak
□ assuming
□ autostism
□ bipaster
□ castanet
□ circumcise
□ clyplaesly
□ crossbill
□ defamable
□ dime
□ dorm
□ elemental
□ enblear
□ flatiron

□ greatcoat
□ heroicomic
□ hould
□ hyplexion
□ inventive
□ javelin
□ jotting
□ leucoin
□ manipular
□ minded
□ opener
□ optimize
□ paubub
□ pigful
□ prizeling

□ psychiatry
□ radiofacy
□ ramate
□ rarely
□ reportage
□ respecting
□ rochead
□ rondel
□ rotatable
□ sabowtra
□ scabby
□ scarfpin
□ seller
□ serviceman
□ sinmersion

□ snuffbox
□ stamper
□ stram
□ syroly
□ tartaric
□ thalamic
□ tradured
□ tumcier
□ turnabout
□ undies
□ urchin
□ wailjoin
□ waybill
□ whicker
□ worrywart

B7. Demographic questionnaire

1. How comfortable are you with *formal statistics* (e.g., chance, confidence intervals & hypothesis test)?
   a. Very uncomfortable
   b. Somewhat uncomfortable
   c. Somewhat comfortable
   d. Very comfortable

2. How comfortable are you with *informal statistics* (e.g., reading tables & graphs of rates & percents)?
   a. Very uncomfortable
   b. Somewhat uncomfortable
   c. Somewhat comfortable
   d. Very comfortable

3. How quantitative is your work, area of study/teaching or daily life? [If retired, use prior occupation.]
   a. Generally non-quantitative (e.g., child care, music, art, English, philosophy)
   b. Minimally quantitative (e.g., business management, education, journalism, health care)
   c. Moderately quantitative (e.g., psychology, sociology, MIS, market research, forecasting)
   d. Highly quantitative (e.g., finance, econometrics, accounting, science, engineering)
   e. Extremely quantitative (e.g., mathematics, statistics)

4. What best describes your occupation? [If retired, you may use your prior occupation.]
   a. Full-time student (not working)
   b. Teacher, elementary/secondary
   c. Teacher, college
   d. Other professions (working full time)
   e. Other (e.g., working part time, homemaker)

5. What best describes your highest level of schooling completed?
   a. Primary school
   b. Secondary school/ High school
   c. Two-year college (associate's degree)
   d. Four year college (bachelor's degree)
   e. Graduate degree (master's or Ph.D)

6. What best describes your fluency in English?
   a. English was a native language in primary school
   b. Became fluent in speaking and reading English after primary school
   c. Not yet fluent in speaking and reading English

7. How many undergraduate &/or graduate statistics courses (e.g., Psych 292) have you completed?

8. How many research method courses (e.g., Psych 291) have you completed?

9. What is your age?

10. What is your program of study and university level?

11. What is your gender?

Appendix C – Value affirmation exercise

Ranking of Personal Characteristics and Values

Below is a list of characteristics and values, some of which may be important to you, some of which may be unimportant. Please rank these values and qualities in order of their importance to you, from 1 to 11 ("1" being the most important item, "11" being the least important). Use each number only once.

____ Artistic skills/Aesthetic appreciation
____ Sense of Humour
____ Relations with friends/Family
____ Spontaneity/Living life in the moment
____ Social Skills
____ Athletics
____ Musical ability/Appreciation
____ Physical attractiveness
____ Creativity
____ Business/Money
____ Romantic values

**On the reverse of the page, the participants were asked to answer two questions, which varied based on their condition:*

If in the <u>value-affirmation</u> condition

1) What was your most important value listed on the previous page?
(the value you ranked number 1)

2) Why do you think this value might be important to you? Describe a time in your life when it has been important. Write as much or as little as you want during this time.

-OR-

If in the <u>control</u> condition:

1) What was your ninth most important value listed on the previous page?
(the value you ranked number 9)

2) Why do you think this value might be important to a typical U of W student? Describe a time in the typical student's life when it may be important. Write as much or as little as you want during this time.

Appendix D – Tables of Results (main, with covariates, and split-groups)

D1. Study 1

Table 1.1

Analysis of Variance – Correct Reasoning Performance – Study1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.53	.529	26.31	.000	.120
Experience	2	.66	.330	16.41	.000	.145
G *Exp	2	.01	.006	.28	.757	.003
Error	193	3.88	.020			
Total	198	5.32				

Table 1.2

Analysis of Variance – Misconceptions Performance – Study1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.20	.203	16.73	.000	.080
Experience	2	.24	.119	9.87	.000	.093
G *Exp	2	.04	.017	1.44	.241	.015
Error	193	2.34	.012			
Total	198	2.88				

Table 1.3

Analysis of Variance – Confidence – Study1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	6.41	6.409	16.91	.000	.081
Experience	2	8.77	4.386	11.57	.000	.107
G *Exp	2	.25	.126	.33	.717	.003
Error	193	73.14	.379			
Total	198	90.83				

Study 1 - Correct Reasoning – List of covariates, in order:

1. PNI
2. CRT
3. WPT
4. Area of Study

Table 1.1.1

Analysis of Covariance (PNI) – Correct Reasoning Performance – Study1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
PNI	1	.29	.287	15.33	.000	.074
Gender	1	.39	.388	20.74	.000	.098
Experience	2	.48	.237	12.69	.000	.117
G *Exp	2	.01	.005	.26	.776	.003
Error	192	3.59	.019			
Total	198	5.32				

Table 1.1.2

Analysis of Covariance (CRT) – Correct Reasoning Performance – Study1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
CRT	1	.29	.287	15.33	.000	.074
Gender	1	.39	.388	20.74	.000	.098
Experience	2	.48	.237	12.69	.000	.117
G *Exp	2	.01	.005	.26	.776	.003
Error	192	3.59	.019			
Total	198	5.32				

Table 1.1.3

Analysis of Covariance (WPT) – Correct Reasoning Performance – Study1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
WPT	1	.88	.882	56.50	.000	.227
Gender	1	.32	.323	20.65	.000	.097
Experience	2	.26	.129	8.25	.000	.079
G *Exp	2	.01	.003	.21	.810	.002
Error	192	3.00	.016			
Total	198	5.32				

Table 1.1.4

Analysis of Covariance (Area of Study) – Correct Reasoning Performance – Study1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Area	1	.20	.201	10.46	.001	.052
Gender	1	.34	.344	17.96	.000	.086
Experience	2	.57	.285	14.85	.000	.134
G *Exp	2	.02	.010	.50	.608	.005
Error	192	3.68	.019			
Total	198	5.32				

Study 1 - Misconceptions – List of covariates, in order:

1. PNI
2. CRT
3. WPT
4. Area of Study

Table 1.2.1

Analysis of Covariance (PNI) – Misconception – Study 1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
PNI	1	.04	.035	2.96	.087	.015
Gender	1	.17	.168	14.05	.000	.068
Experience	2	.19	.097	8.08	.000	.078
G *Exp	2	.03	.017	1.39	.252	.014
Error	192	2.30	.012			
Total	198	2.88				

Table 1.2.2

Analysis of Covariance (CRT) – Misconception – Study 1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
CRT	1	.15	.149	13.11	.000	.064
Gender	1	.08	.077	6.76	.010	.034
Experience	2	.11	.054	4.74	.010	.047
G *Exp	2	.03	.013	1.11	.332	.011
Error	192	2.19	.011			
Total	198	2.88				

Table 1.2.3

Analysis of Covariance (WPT) – Misconception – Study 1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
WPT	1	.02	.019	1.55	.215	.008
Gender	1	.18	.178	14.78	.000	.071
Experience	2	.18	.092	7.62	.001	.074
G *Exp	2	.03	.016	1.33	.267	.014
Error	192	2.32	.012			
Total	198	2.88				

Table 1.2.4

Analysis of Covariance (Area of Study) – Misconception – Study 1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Area	1	.06	.057	4.78	.030	.024
Gender	1	.14	.139	11.72	.001	.058
Experience	2	.21	.104	8.77	.000	.084
G *Exp	2	.04	.021	1.73	.179	.018
Error	192	2.28	.012			
Total	198	2.88				

Study 1 - Confidence – List of covariates, in order:

1. PNI
2. CRT
3. WPT
4. Area of Study

Table 1.3.1

Analysis of Covariance (PNI) – Confidence – Study1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
PNI	1	12.92	12.915	41.17	.000	.177
Gender	1	3.49	3.489	11.12	.001	.055
Experience	2	5.69	2.847	9.08	.000	.086
G *Exp	2	.19	.094	.30	.741	.003
Error	192	60.23	.314			
Total	198	90.83				

Table 1.3.2

Analysis of Covariance (CRT) – Confidence – Study1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
CRT	1	8.28	8.279	24.51	.000	.113
Gender	1	1.70	1.700	5.03	.026	.026
Experience	2	3.95	1.977	5.85	.003	.057
G *Exp	2	.11	.057	.17	.846	.002
Error	192	64.86	.338			
Total	198	90.83				

Table 1.3.3

Analysis of Covariance (WPT) – Confidence – Study1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
WPT	1	3.62	3.619	1.00	.002	.049
Gender	1	4.82	4.819	13.31	.000	.065
Experience	2	5.35	2.68	7.39	.001	.072
G *Exp	2	.19	.095	.26	.771	.003
Error	192	69.52	.362			
Total	198	90.83				

Table 1.3.4

Analysis of Covariance (Area of Study) – Confidence – Study1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Area	1	5.05	5.054	14.25	.000	.069
Gender	1	3.47	3.473	9.79	.002	.049
Experience	2	7.83	3.916	11.04	.000	.103
G *Exp	2	.44	.217	.61	.543	.006
Error	192	68.09	.355			
Total	198	90.83				

D2a. STUDY 2 – Phase 1

Table 2a.1

Analysis of Variance – Correct Reasoning Performance – Study 2 Phase 1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.13	.133	8.46	.009	.308
Condition	1	.00	.001	.04	.853	.002
G *Cond	1	.00	.000	.01	.912	.001
Error	19	.30	.016			
Total	22	.44				

Table 2a.2

Analysis of Variance – Misconceptions Performance – Study 2 Phase 1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.08	.082	5.62	.029	.228
Condition	1	.00	.004	.31	.587	.016
G *Cond	1	.01	.012	.82	.377	.041
Error	19	.28	.015			
Total	22	.39				

Table 2a.3

Analysis of Variance – Confidence – Study 2 Phase 1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.22	.220	.60	.448	.031
Condition	1	.43	.434	1.19	.289	.059
G *Cond	1	.06	.057	.16	.699	.008
Error	19	6.95	.366			
Total	22	8.18				

Study 2 – Phase 1 - Correct Reasoning – List of covariates, in order:

1. Stereotype Endorsement - Stats
2. Stereotype Endorsement - Math
3. Grade 291

Table 2a.1.1

Analysis of Covariance (Stereotype Endorsement- Stats) – Correct Reasoning Performance – Study 2a

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Stats	1	.00	.000	.01	.937	.821
Gender	1	.09	.087	4.40	.041	.077
Error	53	1.05	.020			
Total	55	1.14				

Table 2a.1.2

Analysis of Covariance (Stereotype Endorsement_Math) – Correct Reasoning Performance – Study 2a

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Math	1	.00	.009	.45	.507	.008
Gender	1	.07	.074	3.78	.057	.067
Error	53	1.04	.020			
Total	55	1.14				

Table 2a.1.3

Analysis of Covariance (Grade_291) – Correct Reasoning Performance – Study 2a

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Grade_291	1	.20	.204	12.77	.001	.194
Gender	1	.13	.134	8.38	.006	.136
Error	53	.85	.016			
Total	55	1.14				

Study 2 – Phase 1 - Misconceptions – List of covariates, in order:

1. Stereotype Endorsement - Stats
2. Stereotype Endorsement - Math
3. Grade 291

Table 2a.2.1

Analysis of Covariance (Stereotype Endorsement_Stats) – Misconceptions Performance – Study 2a

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Stats	1	.00	.000	.00	.985	.000
Gender	1	.09	.093	7.35	.009	.122
Error	53	.67	.013			
Total	55	.77				

Table 2a.2.2

Analysis of Covariance (Stereotype Endorsement_Math) – Misconceptions Performance – Study 2a

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Math	1	.00	.001	.076	.784	.001
Gender	1	.09	.087	6.82	.012	.114
Error	53	.67	.013			
Total	55	.76				

Table 2a.1.3

Analysis of Covariance (Grade_291) – Misconceptions Performance – Study 2a

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Grade_291	1	.04	.040	3.39	.071	.060
Gender	1	.11	.113	9.43	.003	.151
Error	53	.63	.012			
Total	55	.77				

Study 2 – Phase 1 - Confidence – List of covariates, in order:

1. Stereotype Endorsement - Stats
2. Stereotype Endorsement - Math
3. Grade 291

Table 2a.3.1

Analysis of Covariance (Stereotype Endorsement_Stats)- Confidence- Study 2 Phase 1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Stats	1	.12	.120	.17	.680	.003
Gender	1	1.77	1.766	2.53	.118	.046
Error	53	36.98	.698			
Total	55	39.01				

Table 2a.3.2

Analysis of Covariance (Stereotype Endorsement_Math)- Confidence- Study 2 Phase 1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Math	1	.00	.001	.08	.784	.001
Gender	1	.09	.087	6.83	.012	.114
Error	53	.67	.013			
Total	55	.76				

Table 2a.3.3

Analysis of Covariance (Grade_291) - Confidence- Study 2 Phase 1

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Grade_291	1	2.12	2.115	3.21	.079	.057
Gender	1	2.55	2.549	3.86	.055	.068
Error	53	34.98	.660			
Total	55	39.02				

D2b. STUDY 2 – Phase 3

Table 2b.1

Analysis of Variance – Correct Reasoning Performance – Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.13	.133	8.46	.009	.308
Condition	1	.00	.001	.04	.853	.002
G *Cond	1	.00	.000	.01	.912	.001
Error	19	.30	.016			
Total	22	.44				

Table 2b.2

Analysis of Variance – Misconceptions Performance – Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.08	.082	5.62	.029	.228
Condition	1	.00	.004	.31	.587	.016
G *Cond	1	.01	.012	.82	.377	.041
Error	19	.28	.015			
Total	22	.39				

Table 2b.3

Analysis of Variance – Confidence Performance – Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.22	.220	.60	.448	.031
Condition	1	.43	.434	1.19	.289	.059
G *Cond	1	.06	.057	.16	.699	.008
Error	19	6.95	.366			
Total	22	8.18				

Study 2 – Phase 3 - Correct Reasoning – List of covariates, in order:

1. PNI
2. NUM
3. Stereotype Endorsement - Stats
4. Stereotype Endorsement - Math
5. Grade 291
6. Word Count

Table 2b.1.1

Analysis of Covariance (PNI)- Correct Reasoning Performance- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
PNI	1	.04	.040	2.80	.112	.135
Gender	1	.15	.151	10.48	.005	.368
Condition	1	.00	.001	.09	.764	.005
G*Cond	1	.01	.005	.35	.559	.019
Error	18	.26	.014			
Total	22	.44				

Table 2b.1.2

Analysis of Covariance (Numeracy)- Correct Reasoning Performance- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Numeracy	1	.01	.013	.79	.385	.042
Gender	1	.09	.087	5.50	.031	.234
Condition	1	.00	.00	.00	.980	.000
G*Cond	1	.00	.001	.08	.783	.004
Error	18	.29	.016			
Total	22	.44				

Table 2b.1.3

Analysis of Covariance (Stereotype Endorsement- Stats)- Correct Reasoning Performance- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Stats	1	.02	.023	1.49	.238	.076
Gender	1	.15	.154	10.07	.005	.359
Condition	1	.00	.001	.05	.825	.003
G*Cond	1	.01	.010	.67	.423	.036
Error	18	.28	.015			
Total	22	.44				

Table 2b.1.4

Analysis of Covariance (Stereotype Endorsement- Math)- Correct Reasoning Performance- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Math	1	.08	.078	6.35	.021	.261
Gender	1	.15	.153	12.44	.002	.409
Condition	1	.00	.002	.15	.700	.008
G *Cond	1	.03	.025	2.06	.169	.103
Error	18	.22	.012			
Total	22	.44				

Table 2b.1.5

Analysis of Covariance (Grade 291)- Correct Reasoning Performance- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Grade_291	1	.06	.059	4.29	.056	.222
Gender	1	.12	.119	8.62	.010	.365
Condition	1	.00	.00	.00	.974	.000
G*Cond	1	.01	.008	.61	.447	.039
Error	15	.21	.016			
Total	19	.39				

Table 2b.1.6

Analysis of Covariance (Word Count)- Correct Reasoning Performance- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Word_Count	1	.09	.088	7.54	.013	.295
Gender	1	.10	.100	8.59	.009	.323
Condition	1	.01	.012	1.03	.323	.054
G*Cond	1	.00	.000	.00	.986	.000
Error	18	.21	.012			
Total	22	.44				

Study 2 – Phase 3 - Misconceptions – List of covariates, in order:

1. PNI
2. NUM
3. Stereotype Endorsement - Stats
4. Stereotype Endorsement - Math
5. Grade 291
6. Word Count

Table 2b.2.1

Analysis of Covariance (PNI)- MISC- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
PNI	1	.06	.058	4.76	.043	.209
Gender	1	.10	.100	8.19	.010	.313
Condition	1	.01	.007	.56	.465	.030
G*Cond	1	.03	.030	2.49	.132	.121
Error	18	.22	.012			
Total	22	.39				

Table 2b.2.2

Analysis of Covariance (Numeracy)- MISC- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Numeracy	1	.00	.003	.19	.669	.010
Gender	1	.06	.060	3.94	.063	.180
Condition	1	.00	.003	.21	.654	.011
G*Cond	1	.01	.014	.91	.354	.048
Error	18	.28	.015			
Total	22	.39				

Table 2b.2.3

Analysis of Covariance (Stereotype Endorsement- Stats)- MISC- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Stats	1	.02	.015	1.03	.324	.054
Gender	1	.10	.096	6.57	.020	.268
Condition	1	.00	.001	.04	.854	.002
G*Cond	1	.03	.026	1.77	.200	.090
Error	18	.26	.015			
Total	22	.39				

Table 2b.2.4

Analysis of Covariance (Stereotype Endorsement- Math)- MISC- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Math	1	.03	.030	2.16	.159	.107
Gender	1	.09	.092	6.64	.019	.269
Condition	1	.00	.001	.04	.841	.002
G*Cond	1	.03	.034	2.45	.135	.120
Error	18	.25	.014			
Total	22	.39				

Table 2b.2.5

Analysis of Covariance (Grade 291)- MISC- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Grade_291	1	.01	.01	.78	.390	.050
Gender	1	.07	.07	5.42	.034	.265
Condition	1	.01	.01	.36	.557	.024
G *Cond	1	.02	.02	1.72	.209	.103
Error	15	.19	.01			
Total	19	.32				

Table 2b.2.6

Analysis of Covariance (Word Count)- MISC- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Word_Count	1	.05	.054	4.30	.053	.193
Gender	1	.06	.062	4.99	.038	.217
Condition	1	.02	.017	1.40	.252	.072
G*Cond	1	.01	.009	.75	.398	.040
Error	18	.23	.012			
Total	22	.39				

Study 2 – Phase 3 - Confidence – List of covariates, in order:

1. PNI
2. NUM
3. Stereotype Endorsement - Stats
4. Stereotype Endorsement - Math
5. Grade 291
6. Word Count

Table 2b.3.1

Analysis of Covariance (PNI)- Confidence- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
PNI	1	.56	.564	1.59	.224	.081
Gender	1	.32	.319	.90	.356	.048
Condition	1	.37	.370	1.04	.320	.055
G *Cond	1	.00	.000	.00	.985	.000
Error	18	6.38	.355			
Total	22	8.18				

Table 2b.3.2

Analysis of Covariance (Numeracy Total)- Confidence- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Numeracy	1	.58	.580	1.64	.217	.084
Gender	1	.02	.022	.06	.806	.003
Condition	1	.61	.614	1.73	.204	.088
G*Cond	1	.01	.008	.02	.884	.001
Error	18	6.37	.354			
Total	22	8.18				

Table 2b.3.3

Analysis of Covariance (Stereotype Endorsement- Stats)- Confidence- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Stats	1	.77	.766	2.23	.153	.110
Gender	1	.50	.495	1.44	.245	.074
Condition	1	.83	.831	2.42	.137	.118
G *Cond	1	.11	.109	.32	.579	.017
Error	18	6.18	.343			
Total	22	8.18				

Table 2b.3.4

Analysis of Covariance (Stereotype Endorsement- Math)- Confidence- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Math	1	.03	.030	2.16	.159	.107
Gender	1	.09	.092	6.64	.019	.269
Condition	1	.00	.001	.04	.841	.002
G *Cond	1	.03	.034	2.45	.135	.120
Error	18	.25	.014			
Total	22	.39				

Table 2b.3.5

Analysis of Covariance (Grade 291)- Confidence- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Grade_291	1	.04	.038	.33	.574	.022
Gender	1	.02	.020	.18	.677	.012
Condition	1	.09	.090	.80	.386	.050
G*Cond	1	.00	.001	.01	.937	.000
Error	15	1.70	.113			
Total	19	1.89				

Table 2b.3.6

Analysis of Covariance (Word Count)- Confidence- Study 2 Phase 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Word_Count	1	.02	.017	.04	.836	.002
Gender	1	.23	.233	.61	.447	.033
Condition	1	.45	.446	1.16	.296	.061
G *Cond	1	.05	.053	.14	.715	.008
Error	18	6.93	.385			
Total	22	8.18				

Split Groups Analyses

Table 2b.1.s

Analysis of Variance – Correct Reasoning – Study 2 Phase 3 – Split Groups

Group	Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Disagree	Gender	1	.03	.034	1.71	.216	.124
	Condition	1	.02	.022	1.07	.321	.082
	G *Cond	0	.00	…	…	…	.000
	Error	12	.24	.020			
	Total	14	.35				
Agree	Gender	1	.00	.001	.58	.482	.103
	Condition	1	.03	.026	11.75	.019	.701
	G *Cond	0	.00	…	…	…	.000
	Error	5	.01	.002			
	Total	7	.09				

Table 2b.2.s

Analysis of Variance – Misconception – Study 2 Phase 3 – Split Groups

Group	Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Disagree	Gender	1	.01	.008	.52	.483	.042
	Condition	1	.05	.046	3.02	.108	.201
	G *Cond	0	.00	…	…	…	.000
	Error	12	.18	.025			
	Total	14	.27				
Agree	Gender	1	.01	.014	.83	.403	.143
	Condition	1	.00	.000	.00	.960	.001
	G *Cond	0	.00	…	…	…	.000
	Error	5	.09	.017			
	Total	7	.12				

Table 2b.3.s

Analysis of Variance – Confidence – Study 2 Phase 3 – Split Groups

Group	Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Disagree	Gender	1	.07	.070	.24	.636	.019
	Condition	1	.17	.166	.56	.470	.044
	G *Cond	0	.00	…	…	…	.000
	Error	12	3.57	.298			
	Total	14	3.75				
Agree	Gender	1	.04	.042	.07	.800	.014
	Condition	1	.83	.833	1.42	.287	.221
	G *Cond	0	.00	…	…	…	.000
	Error	5	2.93	.586			
	Total	7	4.27				

D3. STUDY 3

Table 3.1

Analysis of Variance – Correct Reasoning Performance – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.12	.116	4.60	.034	.031
Condition	1	.04	.042	1.66	.200	.011
Experience	2	.06	.031	1.24	.291	.017
G *Cond	1	.01	.007	.27	.604	.002
G *Exp	2	.16	.080	3.15	.046	.042
Cond *Exp	2	.00	.002	.08	.923	.001
G *Cond *Exp	2	.01	.009	.35	.707	.005
Error	145	3.66	.025			
Total	156	4.21				

Table 3.2

Analysis of Variance – Misconceptions Performance – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.02	.016	1.14	.287	.008
Condition	1	.00	.003	.24	.627	.002
Experience	2	.14	.071	4.98	.008	.064
G *Cond	1	.01	.005	.33	.568	.002
G *Exp	2	.07	.032	2.28	.106	.030
Cond *Exp	2	.00	.002	.14	.867	.002
G *Cond *Exp	2	.00	.002	.16	.854	.002
Error	145	2.06	.014			
Total	156	2.35				

Table 3.1

Analysis of Variance – Confidence – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.02	.023	.05	.830	.000
Condition	1	.55	.545	1.10	.296	.008
Experience	2	.60	.302	.61	.545	.008
G *Cond	1	.07	.070	.14	.709	.001
G *Exp	2	.76	.378	.76	.468	.010
Cond *Exp	2	3.23	1.612	3.25	.042	.043
G *Cond *Exp	2	.54	.269	.54	.582	.007
Error	145	71.93	.496			
Total	156	77.44				

Study 3 - Correct Reasoning – List of covariates, in order:

1. PNI
2. CRT
3. Stereotype Endorsement - Stats
4. Stereotype Endorsement - Math
5. Word Count

Table 3.1.1

Analysis of Covariance (PNI) – Correct Reasoning Performance – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
PNI	1	.30	.304	13.06	.000	.083
Gender	1	.05	.048	2.08	.152	.014
Condition	1	.03	.031	1.32	.252	.009
Experience	2	.03	.015	.64	.527	.009
G *Cond	1	.00	.004	.15	.698	.001
G *Exp	2	.18	.089	3.84	.024	.051
Cond *Exp	2	.00	.002	.07	.929	.001
G *Cond *Exp	2	.00	.002	.09	.914	.001
Error	144	3.35	.023			
Total	156	4.21				

Table 3.1.2

Analysis of Covariance (CRT) – Correct Reasoning Performance – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
CRT	1	.56	.561	26.08	.000	.153
Gender	1	.04	.040	1.86	.174	.013
Condition	1	.02	.022	1.02	.314	.007
Experience	2	.06	.030	1.40	.249	.019
G *Cond	1	.00	.001	.04	.846	.000
G *Exp	2	.08	.041	1.89	.155	.026
Cond *Exp	2	.00	.002	.07	.930	.001
G *Cond *Exp	2	.01	.003	.13	.879	.002
Error	144	3.10	.021			
Total	156	4.21				

Table 3.1.3

Analysis of Covariance (Stereotype Endorsement- Stats) – Correct Reasoning Performance – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Stats	1	.13	.125	5.12	.025	.034
Gender	1	.15	.153	6.25	.014	.042
Condition	1	.03	.029	1.18	.279	.008
Experience	2	.06	.030	1.21	.300	.017
G *Cond	1	.01	.006	.26	.614	.002
G *Exp	2	.17	.082	3.37	.037	.045
Cond *Exp	2	.01	.002	.10	.906	.001
G *Cond *Exp	2	.02	.010	.41	.666	.006
Error	144	3.53	.025			
Total	156	4.21				

Table 3.1.4

Analysis of Covariance (Stereotype Endorsement- Math) – Correct Reasoning Performance – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Math	1	.10	.103	4.17	.043	.028
Gender	1	.15	.149	6.05	.015	.040
Condition	1	.03	.031	1.25	.266	.009
Experience	2	.05	.024	.98	.377	.013
G *Cond	1	.01	.007	.30	.583	.002
G *Exp	2	.15	.076	3.07	.049	.041
Cond *Exp	2	.01	.003	.11	.892	.002
G *Cond *Exp	2	.02	.011	.43	.649	.006
Error	144	3.55	.025			
Total	156	4.21				

Table 3.1.5

Analysis of Covariance (Word Count) – Correct Reasoning Performance – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Word_Count	1	.01	.005	.20	.655	.001
Gender	1	.12	.119	4.70	.032	.032
Condition	1	.04	.041	1.62	.205	.011
Experience	2	.06	.032	1.27	.285	.017
G *Cond	1	.00	.004	.17	.679	.001
G *Exp	2	.16	.081	3.19	.044	.042
Cond *Exp	2	.01	.002	.10	.906	.001
G *Cond *Exp	2	.02	.009	.36	.697	.005
Error	144	3.65	.025			
Total	156	4.21				

Study 3 - Misconceptions – List of covariates, in order:

1. PNI
2. CRT
3. Stereotype Endorsement - Stats
4. Stereotype Endorsement - Math
5. Word Count

Table 3.2.1

Analysis of Covariance (PNI) – Misconception – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
PNI	1	.14	.144	10.86	.001	.070
Gender	1	.00	.002	.16	.688	.001
Condition	1	.00	.001	.11	.741	.001
Experience	2	.07	.035	2.65	.074	.036
G *Cond	1	.00	.003	.20	.652	.001
G *Exp	2	.07	.037	2.76	.067	.037
Cond *Exp	2	.01	.003	.19	.825	.003
G *Cond *Exp	2	.00	.002	.13	.876	.002
Error	144	1.92	.013			
Total	156	4.21				

Table 3.2.2

Analysis of Covariance (CRT) – Misconception – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
CRT	1	.22	.216	16.90	.000	.105
Gender	1	.00	.002	.14	.714	.001
Condition	1	.00	.001	.04	.836	.000
Experience	2	.10	.052	4.03	.020	.053
G *Cond	1	.00	.001	.09	.761	.001
G *Exp	2	.03	.017	1.33	.267	.018
Cond *Exp	2	.01	.002	.19	.831	.003
G *Cond *Exp	2	.01	.004	.33	.717	.005
Error	144	1.84	.013			
Total	156	2.35				

Table 3.2.3

Analysis of Covariance (Stereotype Endorsement- Stats) – Misconception – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Stats	1	.02	.020	1.41	.237	.010
Gender	1	.02	.022	1.54	.216	.011
Condition	1	.00	.002	.14	.709	.001
Experience	2	.14	.070	4.93	.009	.064
G *Cond	1	.00	.004	.32	.576	.002
G *Exp	2	.07	.033	2.34	.100	.031
Cond *Exp	2	.00	.002	.13	.877	.002
G *Cond *Exp	2	.01	.003	.18	.833	.003
Error	144	2.04	.014			
Total	156	2.35				

Table 3.2.4

Analysis of Covariance (Stereotype Endorsement- Math) – Misconception – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Math	1	.02	.021	1.46	.229	.010
Gender	1	.02	.022	1.55	.215	.011
Condition	1	.00	.002	.15	.704	.001
Experience	2	.13	.066	4.67	.011	.061
G *Cond	1	.01	.005	.35	.558	.002
G *Exp	2	.06	.031	2.21	.113	.030
Cond *Exp	2	.00	.002	.12	.887	.002
G *Cond *Exp	2	.01	.003	.20	.822	.003
Error	144	2.04	.014			
Total	156	2.35				

Table 3.2.5

Analysis of Covariance (Word Count) – Misconception – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Word_Count	1	.00	.002	.11	.741	.001
Gender	1	.02	.015	1.07	.302	.007
Condition	1	.00	.003	.24	.623	.002
Experience	2	.14	.068	4.74	.010	.062
G *Cond	1	.01	.006	.39	.532	.003
G *Exp	2	.06	.032	2.23	.112	.030
Cond *Exp	2	.01	.002	.16	.853	.002
G *Cond *Exp	2	.01	.002	.16	.851	.002
Error	144	2.06	.014			
Total	156	2.35				

Study 3 - Confidence – List of covariates, in order:

1. PNI
2. CRT
3. Stereotype Endorsement - Stats
4. Stereotype Endorsement - Math
5. Word Count

Table 3.3.1

Analysis of Covariance (PNI) – Confidence– Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
PNI	1	.67	.673	1.36	.245	.009
Gender	1	.00	.000	.00	.977	.000
Condition	1	.48	.482	.97	.325	.007
Experience	2	.42	.210	.42	.656	.006
G *Cond	1	.05	.052	.11	.745	.001
G *Exp	2	.82	.409	.83	.440	.011
Cond *Exp	2	3.09	1.547	3.13	.047	.042
G *Cond *Exp	2	.49	.247	.50	.609	.007
Error	144	77.26	.495			
Total	156	77.44				

Table 3.3.2

Analysis of Covariance (CRT) – Confidence– Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
CRT	1	.11	.107	.22	.644	.001
Gender	1	.01	.008	.02	.898	.000
Condition	1	.51	.507	1.02	.315	.007
Experience	2	.59	.296	.59	.554	.008
G *Cond	1	.06	.057	.12	.735	.001
G *Exp	2	.67	.333	.67	.515	.009
Cond *Exp	2	3.22	1.609	3.23	.043	.043
G *Cond *Exp	2	.52	.260	.52	.595	.007
Error	144	71.82	.499			
Total	156	77.44				

Table 3.3.3

Analysis of Covariance (Stereotype Endorsement - stats) – Confidence– Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Stats	1	.11	.106	.21	.645	.001
Gender	1	.04	.040	.08	.777	.001
Condition	1	.50	.496	1.00	.320	.007
Experience	2	.60	.297	.60	.552	.008
G *Cond	1	.07	.068	.14	.713	.001
G *Exp	2	.77	.385	.77	.464	.011
Cond *Exp	2	3.24	1.619	3.25	.042	.043
G *Cond *Exp	2	.56	.278	.56	.574	.008
Error	144	71.83	.499			
Total	156	77.44				

Table 3.3.4

Analysis of Covariance (Stereotype Endorsement - math) – Confidence– Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Math	1	.00	.001	.00	.964	.000
Gender	1	.02	.024	.05	.827	.000
Condition	1	.54	.537	1.07	.302	.007
Experience	2	.60	.297	.60	.553	.008
G *Cond	1	.07	.070	.14	.709	.001
G *Exp	2	.75	.377	.76	.472	.010
Cond *Exp	2	3.23	1.613	3.23	.042	.043
G *Cond *Exp	2	.54	.270	.54	.584	.007
Error	144	71.93	.500			
Total	156	77.44				

Table 3.3.5

Analysis of Covariance (Word Count) – Confidence – Study 3

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Word_Count	1	.07	.069	.14	.710	.001
Gender	1	.02	.017	.03	.854	.000
Condition	1	.55	.554	1.11	.294	.008
Experience	2	.60	.299	.60	.551	.008
G *Cond	1	.10	.098	.20	.659	.001
G *Exp	2	.75	.372	.75	.476	.010
Cond *Exp	2	3.11	1.557	3.12	.047	.042
G *Cond *Exp	2	.56	.277	.56	.575	.008
Error	144	71.86	.499			
Total	156	77.44				

Split Groups Analyses

Table 3.1.s

Analysis of Variance – Correct Reasoning – Study 3 – Split Groups

Group	Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Disagree	Gender	1	.07	.071	2.66	.107	.029
	Condition	1	.00	.000	.01	.912	.000
	Experience	2	.07	037	1.38	.257	.030
	G *Cond	1	.00	.002	.07	.798	.001
	G *Exp	2	.10	.048	1.78	.175	.038
	Cond *Exp	2	.01	.007	.25	.779	.006
	G *Cond *Exp	2	.00	.001	.03	.968	.001
	Error	90	2.42	.027			
	Total	101	2.81				
Agree	Gender	1	.05	.049	2.28	.138	.050
	Condition	1	.11	.109	5.07	.029	.105
	Experience	2	.02	.010	.45	.643	.020
	G *Cond	1	.01	.014	.65	.424	.015
	G *Exp	2	.13	.063	2.92	.064	.121
	Cond *Exp	2	.05	.024	1.13	.331	.051
	G *Cond *Exp	2	.06	.027	1.27	.290	.056
	Error	43	.92	.021			
	Total	54	1.34				

Table 3.2.s

Analysis of Variance – Misconception – Study 3 – Split Groups

Group	Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Disagree	Gender	1	.02	.022	1.50	.224	.016
	Condition	1	.00	.002	.16	.691	.002
	Experience	2	.11	055	3.83	.025	.078
	G *Cond	1	.00	.003	.23	.632	.003
	G *Exp	2	.01	.006	.44	.643	.010
	Cond *Exp	2	.04	.017	1.20	.306	.026
	G *Cond *Exp	2	.02	.009	.62	.540	.014
	Error	90	1.29	.014			
	Total	101	1.55				
Agree	Gender	1	.00	.000	.01	.921	.000
	Condition	1	.00	.002	.12	.726	.003
	Experience	2	.03	.014	1.08	.349	.048
	G *Cond	1	.00	.000	.01	.941	.000
	G *Exp	2	.13	.063	4.83	.013	.185
	Cond *Exp	2	.00	.001	.10	.902	.006
	G *Cond *Exp	2	.07	.034	2.61	.085	.108
	Error	43	.56	.013			
	Total	54	.79				

Table 3.3.s

Analysis of Variance – Confidence – Study 3 – Split Groups

Group	Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Disagree	Gender	1	.16	.157	.28	.601	.003
	Condition	1	.57	.569	1.00	.320	.011
	Experience	2	.80	.402	.70	.496	.015
	G *Cond	1	.04	.040	.07	.792	.001
	G *Exp	2	1.49	.742	1.31	.276	.028
	Cond *Exp	2	2.79	1.395	2.45	.092	.052
	G *Cond *Exp	2	.41	.206	.36	.698	.008
	Error	90	51.18	.569			
	Total	101	56.39				
Agree	Gender	1	.04	.035	.08	.773	.002
	Condition	1	.16	.159	.38	.539	.009
	Experience	2	.20	.100	.24	.788	.011
	G *Cond	1	.44	.436	1.05	.311	.024
	G *Exp	2	.02	.009	.02	.978	.002
	Cond *Exp	2	1.59	.793	1.91	.160	.083
	G *Cond *Exp	2	.37	.187	.45	.640	.021
	Error	43	17.84	.415			
	Total	54	20.92				

D4. STUDY 4

Table 4.1

Analysis of Variance – Correct Reasoning Performance– Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.27	.272	18.74	.000	.088
Condition	1	.01	.011	.73	.394	.004
Experience	2	.21	.107	7.38	.001	.071
G *Condition	1	.00	.000	.00	.987	.000
G *Exp	2	.02	.026	1.79	.170	.018
Condition * Exp	2	.00	.002	.148	.863	.002
G *Cond *Exp	2	.04	.021	1.45	.236	.015
Error	193	2.80	.015			
Total	204	3.54				

Table 4.2

Analysis of Variance – Confidence– Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	459.46	459.46	1.27	.261	.007
Condition	1	98.24	98.24	.272	.603	.001
Experience	2	3009.25	1504.62	4.16	.017	.041
G *Cond	1	149.96	149.96	.415	.520	.002
G *Exp	2	258.05	129.02	.357	.700	.004
Cond * Exp	2	613.55	306.77	.848	.430	.009
G *Cond *Exp	2	575.3	287.65	.795	.453	.008
Error	193	69811.84	361.72			
Total	204	75972.46				

Study 4 - Correct Reasoning – List of covariates, in order:

1. PNI
2. CRT
3. Stereotype Endorsement - Stats
4. Stereotype Endorsement - Math
5. Word Count
6. Area of Study

Table 4.1.1

Analysis of Covariance (PNI) – Correct Reasoning – Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
PNI	1	.18	.182	13.22	.000	.064
Gender	1	.10	.103	7.47	.007	.037
Condition	1	.01	.012	.89	.347	.005
Experience	2	.16	.080	5.84	.003	.057
G *Cond	1	.00	.000	.02	.888	.000
G *Exp	2	.04	.018	1.28	.280	.013
Cond *Exp	2	.01	.005	.34	.715	.003
G *Cond *Exp	2	.06	.028	2.00	.139	.020
Error	193	2.66	.014			
Total	205	3.56				

Table 4.1.2

Analysis of Covariance (CRT) – Correct Reasoning – Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
CRT	1	.75	.748	68.99	.000	.263
Gender	1	.04	.044	4.07	.045	.021
Condition	1	.01	.008	.74	.390	.004
Experience	2	.12	.060	5.50	.005	.054
G *Cond	1	.00	.000	.00	.987	.000
G *Exp	2	.04	.021	1.97	.142	.020
Cond *Exp	2	.03	.014	1.33	.268	.014
G *Cond *Exp	2	.03	.016	1.47	.234	.015
Error	193	2.09	.011			
Total	205	3.56				

Table 4.1.3

Analysis of Covariance (Stereotype Endorsement- Stats) – Correct Reasoning – Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Stats	1	.02	.024	1.62	.204	.008
Gender	1	.26	.261	17.90	.000	.085
Condition	1	.01	.014	.97	.326	.005
Experience	2	.21	.105	7.17	.001	.069
G *Cond	1	.00	.001	.06	.804	.000
G *Exp	2	.05	.025	1.69	.187	.017
Cond *Exp	2	.00	.001	.08	.920	.001
G *Cond *Exp	2	.05	.024	1.64	.196	.017
Error	193	2.82	.015			
Total	205	3.56				

Table 4.1.4

Analysis of Covariance (Stereotype Endorsement- Math) – Correct Reasoning – Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
SE_Math	1	.03	.032	2.17	.142	.011
Gender	1	.25	.253	17.36	.000	.083
Condition	1	.02	.015	1.03	.312	.005
Experience	2	.21	.106	7.31	.001	.070
G *Cond	1	.00	.001	.06	.803	.000
G *Exp	2	.05	.023	1.60	.205	.016
Cond *Exp	2	.00	.001	.08	.923	.001
G *Cond *Exp	2	.05	.024	1.68	.190	.017
Error	193	2.81	.015			
Total	205	3.56				

Table 4.1.5

Analysis of Covariance (Word Count) – Correct Reasoning – Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Word_Count	1	.00	.001	.06	.814	.000
Gender	1	.26	.258	17.53	.000	.083
Condition	1	.01	.014	.94	.334	.005
Experience	2	.20	.099	6.75	.001	.065
G *Cond	1	.00	.000	.02	.892	.000
G *Exp	2	.05	.026	1.75	.176	.018
Cond *Exp	2	.00	.001	.09	.916	.001
G *Cond *Exp	2	.04	.020	1.36	.258	.014
Error	193	2.84	.015			
Total	205	3.56				

Table 4.1.6

Analysis of Covariance (Area of Study) – Correct Reasoning – Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Area_of_Study	1	33.97	33.968	.09	.760	.000
Gender	1	528.60	528.604	1.46	.229	.008
Condition	1	77.54	77.539	.21	.644	.001
Experience	2	3226.44	1613.219	4.45	.013	.044
G *Cond	1	124.84	124.835	.34	.558	.002
G *Exp	2	274.45	137.223	.38	.685	.004
Cond *Exp	2	627.54	313.772	.87	.422	.009
G *Cond *Exp	2	531.15	265.575	.73	.482	.008
Error	193	69944.54	362.407			
Total	205	76334.37				

Study 4 - Confidence – List of covariates, in order:

1. PNI
2. CRT
3. Stereotype Endorsement - Stats
4. Stereotype Endorsement - Math
5. Word Count
6. Area of Study

Table 4.2.1

Analysis of Covariance (PNI) – Confidence – Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
PNI	1	2258.41	2258.41	6.44	.012	.032
Gender	1	15.30	15.30	.04	.835	.000
Condition	1	65.57	65.57	.19	.666	.001
Experience	2	2343.53	1171.766	3.34	.038	.033
G*Cond	1	126.62	126.621	.36	.549	.002
G*Exp	2	382.21	191.107	.55	.581	.006
Cond *Exp	2	338.86	169.431	.48	.618	.005
G*Cond *Exp	2	632.83	316.416	.90	.408	.009
Error	193	843425.33	350.881			
Total	205	76334.37				

Table 4.2.2

Analysis of Covariance (CRT) – Confidence– Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
CRT	1	.75	.748	68.99	.000	.263
Gender	1	.04	.044	4.07	.045	.021
Condition	1	.01	.008	.74	.390	.004
Experience	2	.12	.060	5.50	.005	.054
G*Cond	1	.00	.000	.00	.987	.000
G*Exp	2	.04	.021	1.97	.142	.020
Cond *Exp	2	.03	.014	1.33	.268	.014
G*Cond *Exp	2	.03	.016	1.47	.234	.015
Error	193	2.09	.011			
Total	205	3.56				

Table 4.2.3

Analysis of Covariance (Stereotype Endorsement- Stats) – Confidence – Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
ST_Stats	1	.00	.001	.00	.999	.000
Gender	1	505.00	505.006	1.29	.239	.007
Condition	1	80.26	80.256	.22	.639	.001
Experience	2	3192.45	1596.23	4.40	.014	.044
G*Cond	1	126.70	126.699	.35	.555	.002
G*Exp	2	295.63	147.815	.41	.666	.004
Cond *Exp	2	613.94	306.972	.85	.430	.009
G*Cond *Exp	2	536.32	268.160	.74	.479	.008
Error	193	69978.51	362.583			
Total	205	76334.37				

Table 4.2.4

Analysis of Covariance (Stereotype Endorsement- Math) – Confidence – Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
ST_Math	1	64.88	64.880	.18	.673	.001
Gender	1	515.27	515.270	1.42	.234	.007
Condition	1	77.04	77.041	.21	.645	.001
Experience	2	3192.66	1596.330	4.41	.013	.044
G*Cond	1	113.98	113.977	.31	.575	.002
G*Exp	2	310.36	155.179	.43	.652	.004
Cond *Exp	2	580.78	290.392	.80	.450	.008
G*Cond *Exp	2	504.44	252.218	.70	.500	.007
Error	193	69913.63	362.247			
Total	205	76334.37				

Table 4.2.5

Analysis of Covariance (Word Count) – Confidence – Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Word_Count	1	33.97	33.968	.09	.760	.000
Gender	1	528.60	528.604	1.46	.229	.008
Condition	1	77.54	77.539	.21	.644	.001
Experience	2	3226.44	1613.219	4.45	.013	.044
G*Cond	1	124.84	124.835	.34	.558	.002
G*Exp	2	274.45	137.223	.38	.685	.004
Cond *Exp	2	627.54	313.772	.87	.422	.009
G*Cond *Exp	2	531.15	265.575	.73	.482	.008
Error	193	69944.54	362.407			
Total	205	76334.37				

Table 4.2.6

Analysis of Covariance (Area of Study) – Correct Reasoning – Study 4

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Area	1	.09	.085	5.96	.016	.030
Gender	1	.16	.162	11.31	.001	.055
Condition	1	.03	.025	1.73	.191	.009
Experience	2	.13	.065	4.52	.012	.045
G*Cond	1	.00	.001	.055	.815	.000
G*Exp	2	.06	.029	2.02	.135	.021
Cond *Exp	2	.00	.002	.128	.880	.001
G*Cond *Exp	2	.06	.032	2.21	.112	.022
Error	193	2.76	.014			
Total	205	3.56				

D5. Study 5

Table 5.1

Analysis of Variance – Quantitative: Number of Sets – Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	17.33	17.334	3.028	.083	.014
Experience	2	9.55	4.776	.834	.436	.008
Gender * Exp	2	11.41	5.705	.997	.371	.009
Error	213	1219.33	5.725			
Total	218	1258.56	17.334			

Table 5.2

Analysis of Variance – Quantitative: Deviations from Ideal – Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.20	.199	2.90	.090	.013
Experience	2	.17	.084	1.23	.295	.011
G * Exp	2	.11	.053	.77	.463	.007
Error	213	14.60	.069			
Total	218	15.08				

Table 5.3

Analysis of Variance – Qualitative: Quality of Rationales – Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	1.27	1.266	2.97	.086	.014
Experience	2	3.22	1.607	3.77	.025	.034
G * Exp	2	.78	.390	.92	.402	.009
Error	213	90.77	.426			
Total	218	95.72				

Study 5 - Number of Sets– List of covariates, in order:

1. Area of Study
2. Level of Comfort- Formal Stats
3. Level of Comfort- Informal Stats

Table 5.1

Analysis of Covariance (Area of Study) –Number of Sets- Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Area	1	.02	.017	.25	.619	.001
Gender	1	.16	.160	2.33	.128	.011
Experience	2	.16	.080	1.16	.315	.011
G *Exp	2	14.58	.056	.81	.446	.008
Error	212	59.02	.069			
Total	218	15.08				

Table 5.2

Analysis of Covariance (Comfort - Formal) –Number of Sets- Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Comfort - Formal	1	.02	.017	.25	.619	.001
Gender	1	.16	.160	2.33	.128	.011
Experience	2	.16	.080	1.16	.315	.011
G *Exp	2	14.58	.056	.81	.446	.008
Error	212	59.02	.069			
Total	218	15.08				

Table 5.3

Analysis of Covariance (Comfort - Informal) –Number of Sets- Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Comfort - Informal	1	.02	.017	.25	.619	.001
Gender	1	.16	.160	2.33	.128	.011
Experience	2	.16	.080	1.16	.315	.011
G *Exp	2	14.58	.056	.81	.446	.008
Error	212	59.02	.069			
Total	218	15.08				

Study 5 - Deviations from Ideal– List of covariates, in order:

1. Number of Sets
2. Area of Study
3. Level of Comfort- Formal Stats
4. Level of Comfort- Informal Stats

Table 5.2.1

Analysis of Covariance (Number of Sets) – Deviations from Ideal – Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Number_Sets	1	9.92	9.922	450.00	.000	.680
Gender	1	.01	.005	.22	.639	.001
Experience	2	.02	.011	.51	.603	.005
G *Exp	2	.00	.001	.03	.974	.000
Error	212	4.68	.022			
Total	218	15.08				

Table 5.2.2

Analysis of Covariance (Area of Study) – Deviations from Ideal – Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Area	1	.02	.017	.25	.619	.001
Gender	1	.16	.160	2.33	.128	.011
Experience	2	.16	.080	1.16	.315	.011
G *Exp	2	14.58	.056	.81	.446	.008
Error	212	59.02	.069			
Total	218	15.08				

Table 5.2.3

Analysis of Covariance (Level of Comfort- Formal Stats) – Deviations from Ideal – Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Comfort_Formal	1	.14	.144	2.12	.147	.010
Gender	1	.20	.204	2.99	.085	.014
Experience	2	.15	.075	1.11	.332	.010
G *Exp	2	.11	.055	.81	.447	.008
Error	212	14.45	.068			
Total	218	15.08				

Table 5.2.4

Analysis of Covariance (Level of Comfort- Informal Stats) – Deviations from Ideal – Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Comfort_Informal	1	.41	.405	6.05	.015	.028
Gender	1	.25	.246	3.68	.057	.017
Experience	2	.14	.071	1.06	.347	.010
G *Exp	2	.09	.047	.70	.498	.007
Error	212	14.19	.067			
Total	218	15.08				

Study 5 - Quality of Rationales – List of covariates, in order:

1. Number of Sets
2. Area of Study
3. Level of Comfort- Formal Stats
4. Level of Comfort- Informal Stats

Table 5.3.1

Analysis of Covariance (Number of Sets) – Quality of Rationales – Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Number_of_Sets	1	2.16	2.162	5.17	.024	.024
Gender	1	1.32	1.316	3.15	.077	.015
Experience	2	2.70	1.348	3.23	.042	.030
G *Exp	2	.68	.340	.81	.444	.008
Error	212	88.61	.418			
Total	218	95.72				

Table 5.3.2

Analysis of Covariance (Area of Study) – Quality of Rationales – Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Area	1	1.02	1.021	2.41	.122	.011
Gender	1	.71	.714	1.69	.195	.008
Experience	2	2.93	1.466	3.46	.033	.032
G *Exp	2	.63	.317	.75	.474	.007
Error	212	89.75	.423			
Total	218	95.72				

Table 5.3.3

Analysis of Covariance (Level of Comfort- Formal Stats) – Quality of Rationales – Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Comfort_Formal	1	2.04	2.037	4.87	.028	.022
Gender	1	1.53	1.532	3.66	.057	.017
Experience	2	2.95	1.476	3.53	.031	.032
G *Exp	2	.82	.409	.98	.378	.009
Error	212	88.73	.419			
Total	218	95.72				

Table 5.3.4

Analysis of Covariance (Level of Comfort- Informal Stats) – Quality of Rationales – Study 5

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Comfort_Informal	1	22.39	22.386	69.40	.000	.247
Gender	1	.31	.310	.96	.328	.005
Experience	2	2.03	1.015	3.15	.045	.029
G *Exp	2	1.23	.613	1.90	.152	.018
Error	212	68.38	.323			
Total	218	95.72				

D6. STUDY 6

Table 6.1

Analysis of Variance – Quantitative: Deviations from Ideal – Study 6

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.02	.024	.36	.550	.002
Experience	2	.19	.092	1.41	.248	.014
G *Exp	2	.14	.071	1.09	.340	.011
Error	202	13.27	.066			
Total	207	13.64				

Table 6.2

Analysis of Variance – Qualitative: Quality of Rationales – Study 6

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	2.89	2.894	6.59	.011	.032
Experience	2	.14	.069	.16	.855	.002
G *Exp	2	.64	.318	.73	.486	.007
Error	202	88.66	.439			
Total	207	92.06				

Study 6 - Deviations from ideal– List of covariates, in order:

1. Area of Study
2. Level of Comfort- Formal Stats
3. Level of Comfort- Informal Stats

Table 6.1.1

Analysis of Covariance (Area of Study) – Deviations from ideal – Study 6

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Area	1	.11	.105	1.62	.205	.008
Gender	1	.06	.064	.98	.323	.005
Experience	2	.15	.076	1.17	.314	.012
G *Exp	2	.16	.081	1.25	.289	.012
Error	200	13.02	.065			
Total	206	13.56				

Table 6.1.2

Analysis of Covariance (Level of Comfort- Formal Stats) – Deviations from ideal – Study 6

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Comfort_Formal	1	.00	.000	.00	.961	.000
Gender	1	.03	.126	.39	.533	.002
Experience	2	.20	.102	1.54	.216	.015
G *Exp	2	.17	.083	1.26	.285	.013
Error	199	13.09	.066			
Total	205	13.51				

Table 6.1.3

Analysis of Covariance (Level of Comfort- Informal Stats) – Deviations from ideal – Study 6

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Comfort_Informal	1	.03	.034	.53	.470	.003
Gender	1	.03	.032	.48	.488	.002
Experience	2	.20	.100	1.53	.220	.015
G *Exp	2	.18	.088	1.35	.262	.013
Error	200	13.09	.065			
Total	206	13.56				

Study 6 - Quality of Rationales – List of covariates, in order:

1. Area of Study
2. Level of Comfort- Formal Stats
3. Level of Comfort- Informal Stats

Table 6.2.1

Analysis of Covariance (Area of Study) – Quality of Rationales – Study 6

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Area_of_Study	1	1.90	1.899	4.38	.038	.021
Gender	1	4.03	4.034	9.30	.003	.044
Experience	2	.09	.046	.11	.898	.001
G *Exp	2	.55	.277	.64	.529	.006
Error	200	86.73	.434			
Total	206	92.05				

Table 6.2.2

Analysis of Covariance (Level of Comfort- Formal Stats) – Quality of Rationales – Study 6

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Area_of_Study	1	1.90	1.899	4.38	.038	.021
Gender	1	4.03	4.034	9.30	.003	.044
Experience	2	.09	.046	.11	.898	.001
G *Exp	2	.55	.277	.64	.529	.006
Error	200	86.73	.434			
Total	206	92.05				

Table 6.2.3

Analysis of Covariance (Level of Comfort- Informal Stats) – Quality of Rationales – Study 6

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Area_of_Study	1	1.90	1.899	4.38	.038	.021
Gender	1	4.03	4.034	9.30	.003	.044
Experience	2	.09	.046	.11	.898	.001
G *Exp	2	.55	.277	.64	.529	.006
Error	200	86.73	.434			
Total	206	92.05				

D7. STUDIES 5 & 6

Table 7.1

Analysis of Variance – Quantitative: Deviations from Ideal – Studies 5 & 6

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.05	.046	.68	.410	.002
Experience	2	.34	.172	2.56	.078	.012
FORMAT	1	.03	.030	.45	.501	.001
Gender* Exp	2	.18	.088	1.31	.270	.006
Gender*Format	1	.18	.182	2.71	.100	.007
Exp * Format	2	.00	.001	.02	.985	.000
G * Exp * Format	2	.06	.027	.41	.667	.002
Error	415	27.86	.067			
Total	426	28.75				

Table 7.2

Analysis of Variance – Qualitative: Quality of Rationales – Studies 5 & 6

Source	*Df*	*SS*	*MS*	*F*	*p*	η^2_p
Gender	1	.14	.139	.32	.570	.001
Experience	2	2.08	1.039	2.40	.092	.011
FORMAT	1	5.75	5.746	13.29	.000	.031
Gender* Exp	2	.048	.024	.055	.947	.000
Gender*Format	1	3.97	3.966	9.17	.003	.022
Exp * Format	2	1.55	.773	1.79	.169	.009
G * Exp * Format	2	1.37	.683	1.58	.207	.008
Error	415	179.43	.432			
Total	426	196.68				

Studies 5 & 6 - Deviations from ideal– List of covariates, in order:

1. Area of Study
2. Level of Comfort- Formal Stats
3. Level of Comfort- Informal Stats

Table 7.1.1

Analysis of Covariance (Area of Study) – Deviations from ideal – Studies 5 & 6

Source	*Df*	*SS*	*MS*	*F*	*p*	η^2_p
Area	1	.10	.099	1.48	.224	.004
Gender	1	.01	.013	.19	.662	.000
Experience	2	.32	.158	2.37	.095	.011
FORMAT	1	.03	.033	.49	.486	.001
Gender* Exp	2	.19	.096	1.44	.238	.007
Gender*Format	1	.19	.193	2.89	.090	.007
Exp * Format	2	.00	.001	.02	.985	.000
G * Exp * Format	2	.07	.034	.51	.604	.002
Error	413	27.62	.067			
Total	425	28.66				

Table 7.1.2
Analysis of Covariance (Level of Comfort- Formal Stats) – Deviations from ideal – Studies 5 & 6

Source	*Df*	*SS*	*MS*	*F*	*p*	η^2_p
Comfort_Formal	1	.07	.067	1.01	.317	.002
Gender	1	.04	.041	.61	.434	.001
Experience	2	.35	.172	2.57	.078	.012
FORMAT	1	.03	.031	.46	.499	.001
Gender* Exp	2	.21	.104	1.55	.214	.007
Gender*Format	1	.19	.194	2.90	.089	.007
Exp * Format	2	.00	.002	.03	.967	.000
G * Exp * Format	2	.06	.030	.45	.638	.002
Error	412	27.62	.067			
Total	424	28.62				

Table 7.1.3
Analysis of Covariance (Level of Comfort- Informal Stats) – Deviations from ideal – Studies 5 & 6

Source	*Df*	*SS*	*MS*	*F*	*p*	η^2_p
Comfort_Informal	1	.36	.363	5.48	.020	.013
Gender	1	.05	.049	.74	.391	.002
Experience	2	.33	.162	.45	.087	.012
FORMAT	1	.04	.041	.62	.431	.002
Gender* Exp	2	.20	.097	.47	.232	.007
Gender*Format	1	.22	.223	.37	.067	.008
Exp * Format	2	.01	.002	.04	.964	.000
G * Exp * Format	2	.08	.037	.56	.569	.003
Error	413	27.36	.066			
Total	425	28.66				

<u>Studies 5 & 6 - Quality of Rationales – List of covariates, in order:</u>

1. Area of Study
2. Level of Comfort- Formal Stats
3. Level of Comfort- Informal Stats

Table 7.2.1
Analysis of Covariance (Area of Study) – Quality of Rationales – Study 6

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Area_of_Study	1	2.80	2.805	6.56	.011	.016
Gender	1	.62	.623	1.46	.228	.004
Experience	2	1.47	.738	1.73	.179	.008
FORMAT	1	5.92	5.917	13.84	.000	.032
Gender* Exp	2	.04	.020	.05	.955	.000
Gender*Format	1	3.94	3.940	9.21	.003	.022
Exp * Format	2	1.71	.856	2.00	.136	.010
G * Exp * Format	2	1.23	.563	1.32	.269	.006
Error	413	176.59	.428			
Total	425	196.63				

Table 6.2.2

Analysis of Covariance (Level of Comfort- Formal Stats) – Quality of Rationales – Study 6

Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Comfort_Formal	1	.43	.431	1.00	.319	.002
Gender	1	.12	.119	.28	.599	.001
Experience	2	1.92	.959	2.22	.110	.011
FORMAT	1	5.69	5.694	13.18	.000	.031
Gender* Exp	2	.02	.010	.02	.978	.000
Gender*Format	1	3.88	3.878	8.98	.003	.021
Exp * Format	2	1.46	.73	1.69	.186	.008
G * Exp * Format	2	1.36	.679	1.57	.209	.008
Error	412	178.02	.432			
Total	424	195.15				

Table 6.2.3

Analysis of Covariance (Level of Comfort- Informal Stats) – Quality of Rationales – Study 6

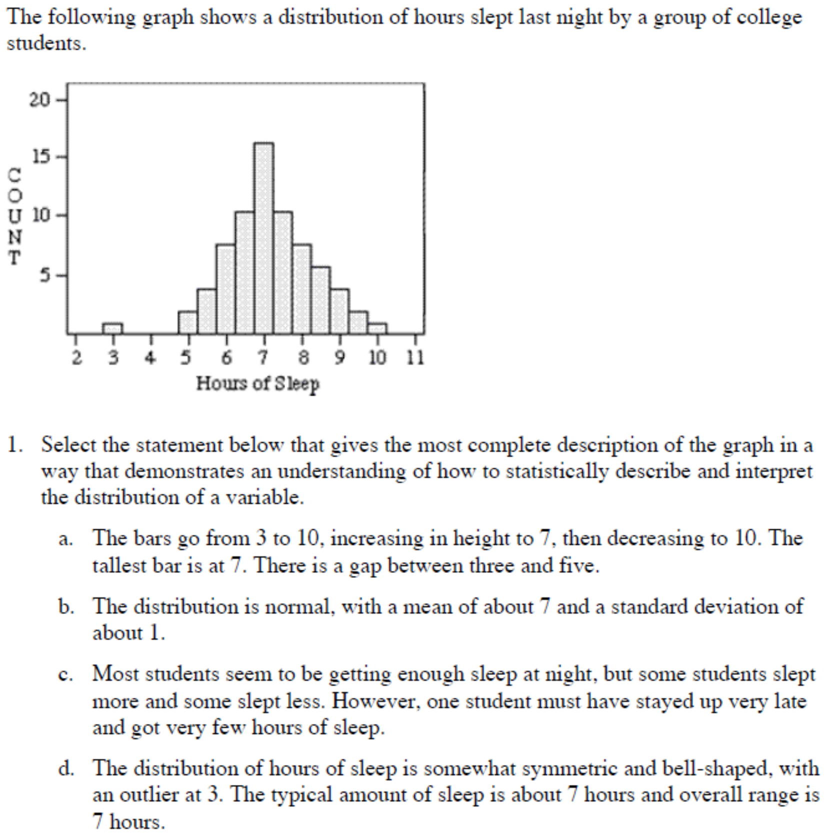
Source	*df*	*SS*	*MS*	*F*	*p*	η^2_p
Comfort_Informal	1	1.30	1.296	3.01	.084	.007
Gender	1	.11	.114	.27	.607	.001
Experience	2	1.94	.967	2.24	.107	.011
FORMAT	1	5.58	5.577	12.93	.000	.030
Gender* Exp	2	.08	.041	.10	.910	.000
Gender*Format	1	4.22	4.217	9.78	.002	.023
Exp * Format	2	1.48	.738	1.71	.182	.008
G * Exp * Format	2	1.21	.603	1.40	.248	.007
Error	413	178.10	.431			
Total	425					

Appendix E – CAOS (including learning outcome for each problem)

E1. CAOS Questions and Learning Outcomes

QUESTION 1

Learning outcome: *Ability to describe and interpret the overall distribution of a variable as displayed in a histogram, including referring to the context of the data.*



The following graph shows a distribution of hours slept last night by a group of college students.

1. Select the statement below that gives the most complete description of the graph in a way that demonstrates an understanding of how to statistically describe and interpret the distribution of a variable.

    a. The bars go from 3 to 10, increasing in height to 7, then decreasing to 10. The tallest bar is at 7. There is a gap between three and five.

    b. The distribution is normal, with a mean of about 7 and a standard deviation of about 1.

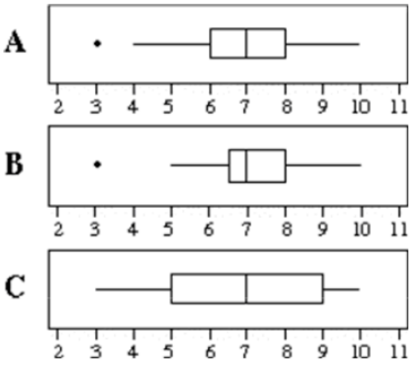
    c. Most students seem to be getting enough sleep at night, but some students slept more and some slept less. However, one student must have stayed up very late and got very few hours of sleep.

    d. The distribution of hours of sleep is somewhat symmetric and bell-shaped, with an outlier at 3. The typical amount of sleep is about 7 hours and overall range is 7 hours.

QUESTION 2

Learning Outcome: *Ability to recognize two different graphical representations of the same data (boxplot and histogram.*

2. Which box plot seems to be graphing the same data as the histogram in question 1?

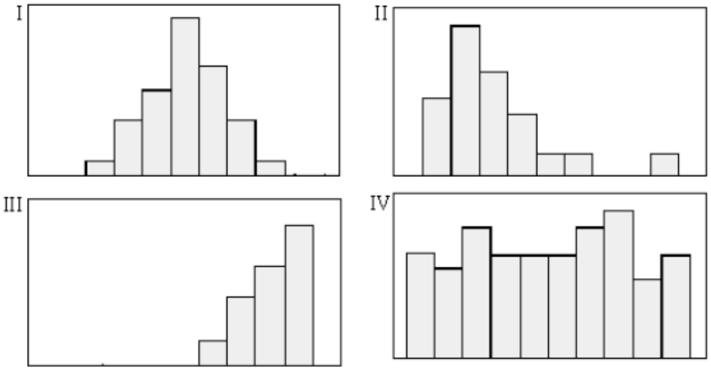


a. Boxplot A.
b. Boxplot B.
c. Boxplot C.

QUESTION 3

Learning Outcomes: *Ability to visualize and match a histogram to a description of a variable (negatively skewed distribution for scores on an easy quiz).*

Four histograms are displayed below. For each item, match the description to the appropriate histogram.



3. A distribution for a set of quiz scores where the quiz was very easy is represented by:
   a. Histogram I.
   b. Histogram II.
   c. Histogram III.
   d. Histogram IV.

QUESTION 4

Learning Outcome: *Ability to visualize and match a histogram to a description of a variable (bell-shaped distribution for wrist circumferences of newborn female infants).*

4. A distribution for a set of wrist circumferences (measured in centimeters) taken from the right wrist of a random sample of newborn female infants is represented by:

   a. Histogram I.

   b. Histogram II.

   c. Histogram III.

   d. Histogram IV.

QUESTION 5

Learning Outcome: *Ability to visualize and match a histogram to a description of a variable (uniform distribution for the last digit of phone numbers sampled from a phone book)*

5. A distribution for the last digit of phone numbers sampled from a phone book (i.e., for the phone number 968-9667, the last digit, 7, would be selected) is represented by:

   a. Histogram I.

   b. Histogram II.

   c. Histogram III.

   d. Histogram IV.

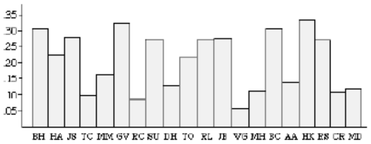
QUESTION 6

Learning Outcome: *Understanding that to properly describe the distribution (shape, center, and spread) of a quantitative variable, a graph like a histogram as needed.*

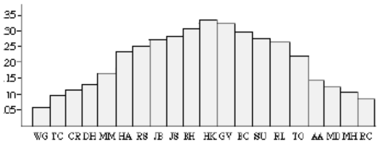
6. A baseball fan likes to keep track of statistics for the local high school baseball team. One of the statistics she recorded is the proportion of hits obtained by each player based on the number of times at bat as shown in the table below. Which of the following graphs gives the best display of the distribution of proportion of hits in that it allows the baseball fan to describe the shape, center and spread of the variable, proportion of hits?

Player	Proportion of hits	Player	Proportion of hits	Player	Proportion of hits
BH	0.305	SU	0.270	BC	0.301
HA	0.229	DH	0.136	AA	0.143
JS	0.281	TO	0.218	HK	0.341
TC	0.097	RL	0.267	RS	0.261
MM	0.167	JB	0.270	CR	0.115
GV	0.333	WG	0.054	MD	0.125
RC	0.085	MH	0.108		

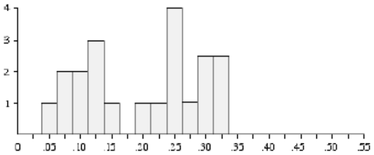
A



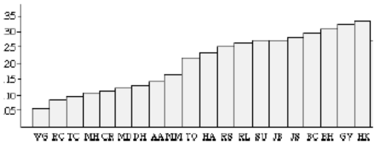
B



C

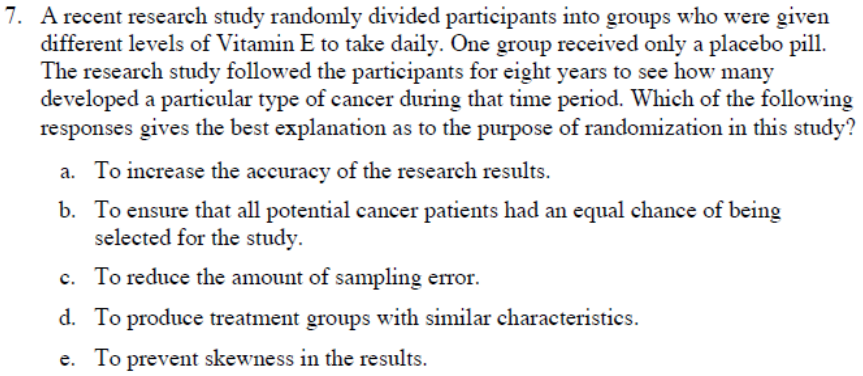


D



QUESTION 7

Learning Outcome: *Understanding of the purpose of randomization in an experiment.*



7. A recent research study randomly divided participants into groups who were given different levels of Vitamin E to take daily. One group received only a placebo pill. The research study followed the participants for eight years to see how many developed a particular type of cancer during that time period. Which of the following responses gives the best explanation as to the purpose of randomization in this study?

   a. To increase the accuracy of the research results.

   b. To ensure that all potential cancer patients had an equal chance of being selected for the study.

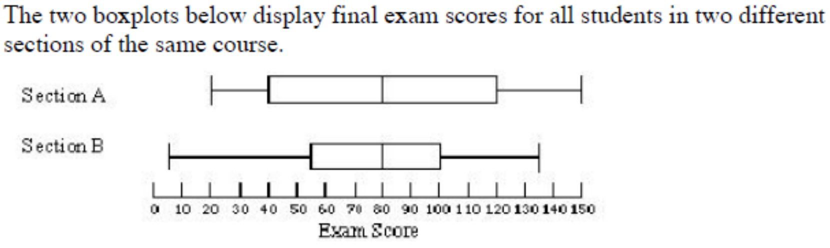
   c. To reduce the amount of sampling error.

   d. To produce treatment groups with similar characteristics.

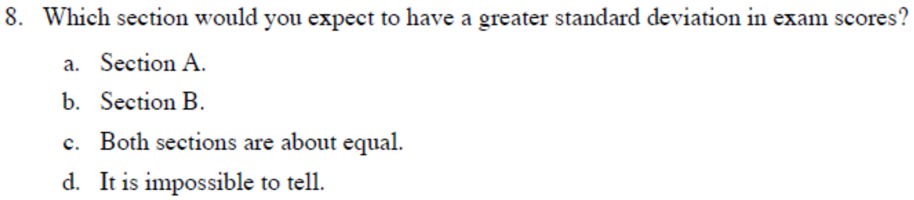
   e. To prevent skewness in the results.

QUESTION 8

Learning Outcome: *Ability to determine which of two boxplots represents a larger standard deviation.*



The two boxplots below display final exam scores for all students in two different sections of the same course.



8. Which section would you expect to have a greater standard deviation in exam scores?

   a. Section A.

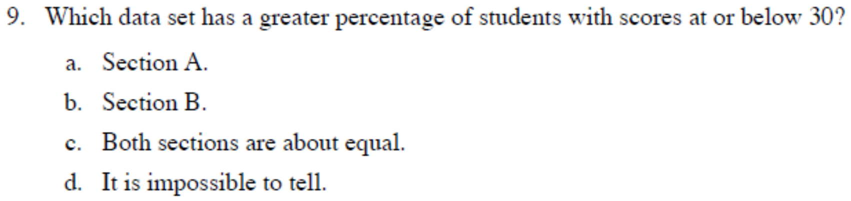
   b. Section B.

   c. Both sections are about equal.

   d. It is impossible to tell.

QUESTION 9

Learning Outcome: *Understanding that boxplots do not provide accurate estimates for percentages of data above or below values except for the quartiles.*



9. Which data set has a greater percentage of students with scores at or below 30?

   a. Section A.

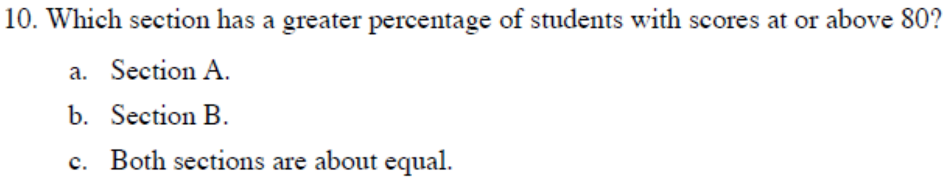
   b. Section B.

   c. Both sections are about equal.

   d. It is impossible to tell.

QUESTION 10

Learning Outcome: *Understanding of the interpretation of a median in the context of boxplots.*



10. Which section has a greater percentage of students with scores at or above 80?

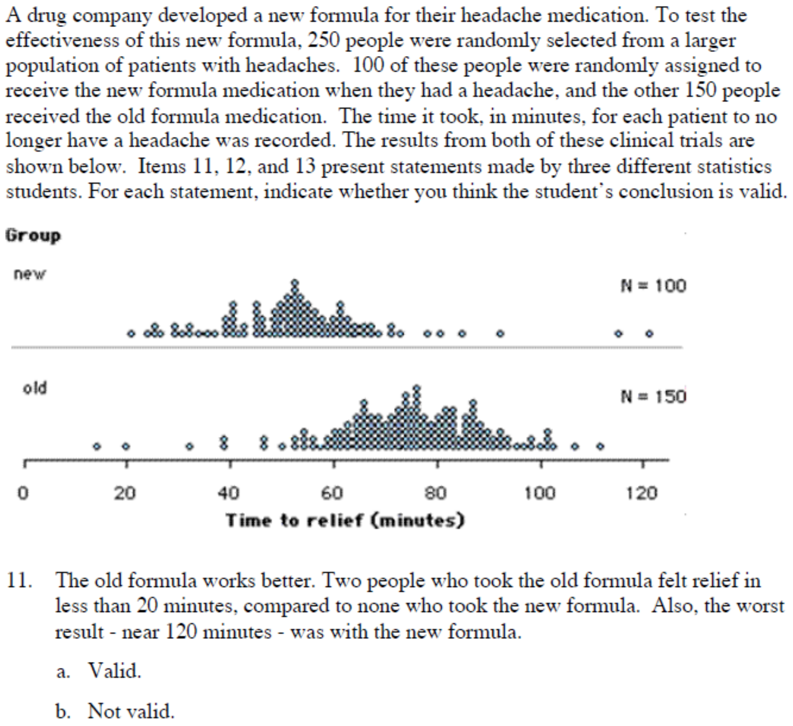
a. Section A.

b. Section B.

c. Both sections are about equal.

QUESTION 11

Learning Outcome: *Ability to compare groups by considering where most of the data are, and focusing on distribution as single entities.*



A drug company developed a new formula for their headache medication. To test the effectiveness of this new formula, 250 people were randomly selected from a larger population of patients with headaches. 100 of these people were randomly assigned to receive the new formula medication when they had a headache, and the other 150 people received the old formula medication. The time it took, in minutes, for each patient to no longer have a headache was recorded. The results from both of these clinical trials are shown below. Items 11, 12, and 13 present statements made by three different statistics students. For each statement, indicate whether you think the student's conclusion is valid.

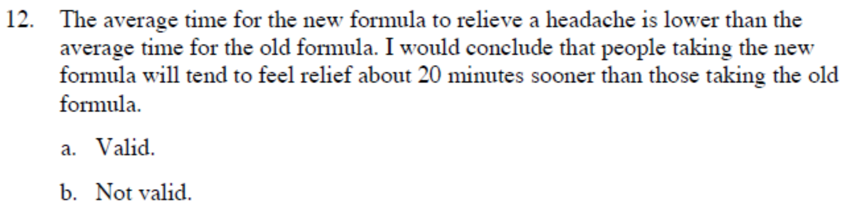
11. The old formula works better. Two people who took the old formula felt relief in less than 20 minutes, compared to none who took the new formula. Also, the worst result - near 120 minutes - was with the new formula.

a. Valid.

b. Not valid.

QUESTION 12

Learning Outcome: *Ability to compare groups by comparing differences in averages.*



12. The average time for the new formula to relieve a headache is lower than the average time for the old formula. I would conclude that people taking the new formula will tend to feel relief about 20 minutes sooner than those taking the old formula.

a. Valid.

b. Not valid.

QUESTION 13

Learning Outcome: *Understanding that comparing two groups does not require equal sample sizes in each group, especially if both sets of data are large.*

13. I would not conclude anything from these data. The number of patients in the two groups is not the same so there is no fair way to compare the two formulas.

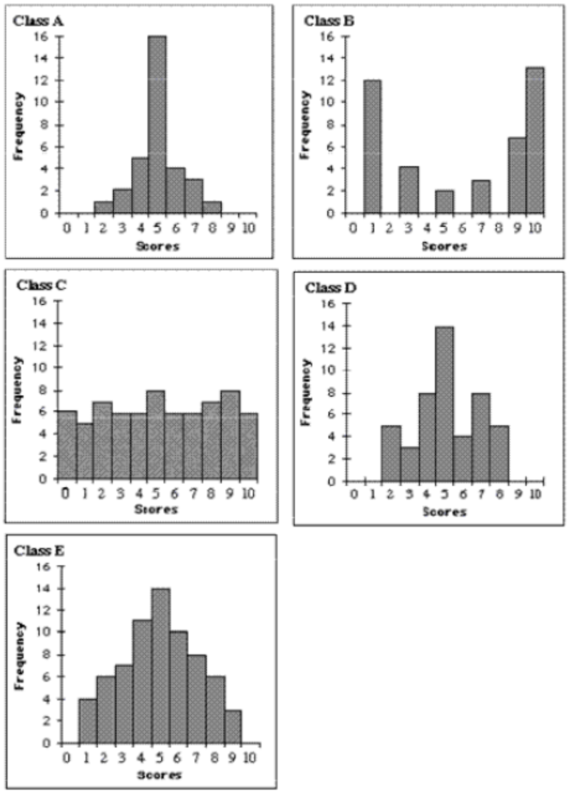
    a. Valid.

    b. Not valid.

QUESTION 14

Learning Outcome: *Ability to correctly estimate and compare standard deviations for different histograms. Understands lowest standard deviation would be for a graph with the least spread (typically) away from the center.*

Five histograms are presented below. Each histogram displays test scores on a scale of 0 to 10 for one of five different statistics classes.



14. Which of the classes would you expect to have the lowest standard deviation, and why?

    a. Class A, because it has the most values close to the mean.

    b. Class B, because it has the smallest number of distinct scores.

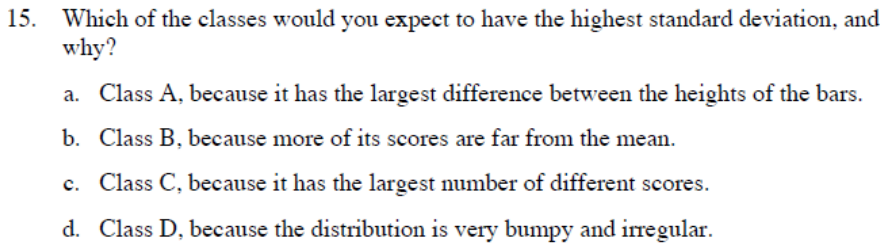
    c. Class C, because there is no change in scores.

    d. Class A and Class D, because they both have the smallest range.

    e. Class E, because it looks the most normal.

QUESTION 15

Learning Outcome: *Ability to correctly estimate standard deviation for different histograms. Understands highest standard deviation would be for a graph with the most spread (typically) away from the center.*



15. Which of the classes would you expect to have the highest standard deviation, and why?

    a. Class A, because it has the largest difference between the heights of the bars.

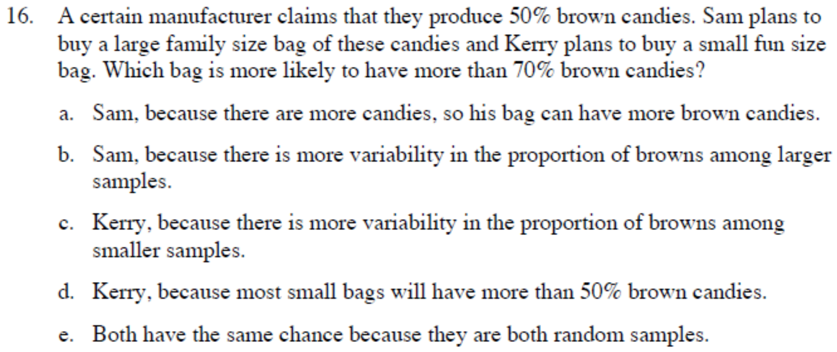
    b. Class B, because more of its scores are far from the mean.

    c. Class C, because it has the largest number of different scores.

    d. Class D, because the distribution is very bumpy and irregular.

QUESTION 16

Learning Outcome*: Understanding the statistics from small samples vary more than statistics from large samples.*



16. A certain manufacturer claims that they produce 50% brown candies. Sam plans to buy a large family size bag of these candies and Kerry plans to buy a small fun size bag. Which bag is more likely to have more than 70% brown candies?

    a. Sam, because there are more candies, so his bag can have more brown candies.

    b. Sam, because there is more variability in the proportion of browns among larger samples.

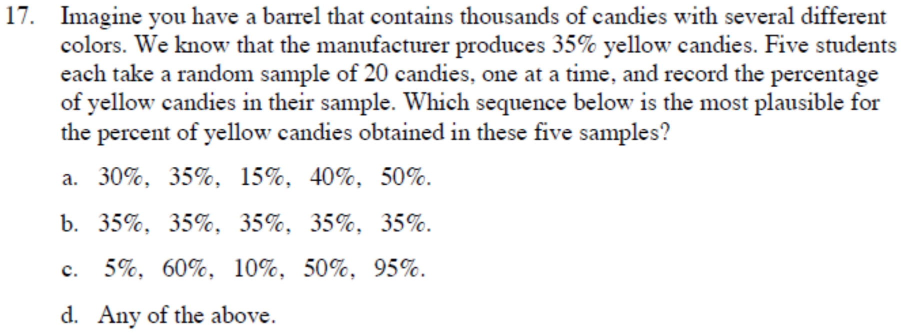
    c. Kerry, because there is more variability in the proportion of browns among smaller samples.

    d. Kerry, because most small bags will have more than 50% brown candies.

    e. Both have the same chance because they are both random samples.

QUESTION 17

Learning Outcome: *Understanding of expected patterns in sampling variability.*



17. Imagine you have a barrel that contains thousands of candies with several different colors. We know that the manufacturer produces 35% yellow candies. Five students each take a random sample of 20 candies, one at a time, and record the percentage of yellow candies in their sample. Which sequence below is the most plausible for the percent of yellow candies obtained in these five samples?

    a. 30%, 35%, 15%, 40%, 50%.

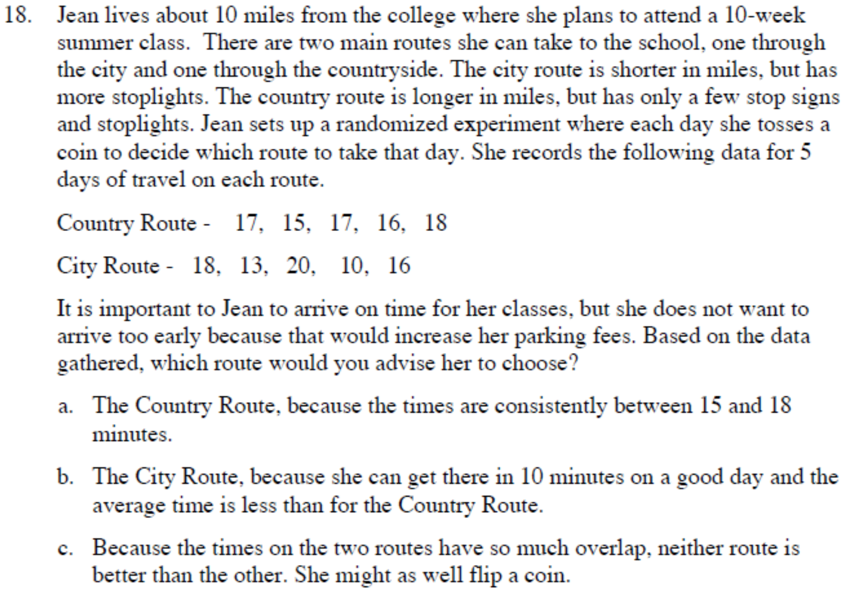
    b. 35%, 35%, 35%, 35%, 35%.

    c. 5%, 60%, 10%, 50%, 95%.

    d. Any of the above.

QUESTION 18

Learning Outcome: *Understanding of the meaning of variability in the context of repeated measurements, and in a context where small variability is desired.*



18. Jean lives about 10 miles from the college where she plans to attend a 10-week summer class. There are two main routes she can take to the school, one through the city and one through the countryside. The city route is shorter in miles, but has more stoplights. The country route is longer in miles, but has only a few stop signs and stoplights. Jean sets up a randomized experiment where each day she tosses a coin to decide which route to take that day. She records the following data for 5 days of travel on each route.

    Country Route - 17, 15, 17, 16, 18

    City Route - 18, 13, 20, 10, 16

    It is important to Jean to arrive on time for her classes, but she does not want to arrive too early because that would increase her parking fees. Based on the data gathered, which route would you advise her to choose?

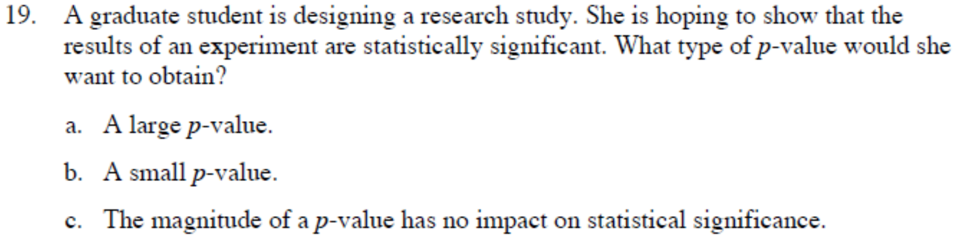
    a. The Country Route, because the times are consistently between 15 and 18 minutes.

    b. The City Route, because she can get there in 10 minutes on a good day and the average time is less than for the Country Route.

    c. Because the times on the two routes have so much overlap, neither route is better than the other. She might as well flip a coin.

QUESTION 19

Learning Outcome: *Understanding that low p-values are desirable in research studies.*



19. A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of *p*-value would she want to obtain?

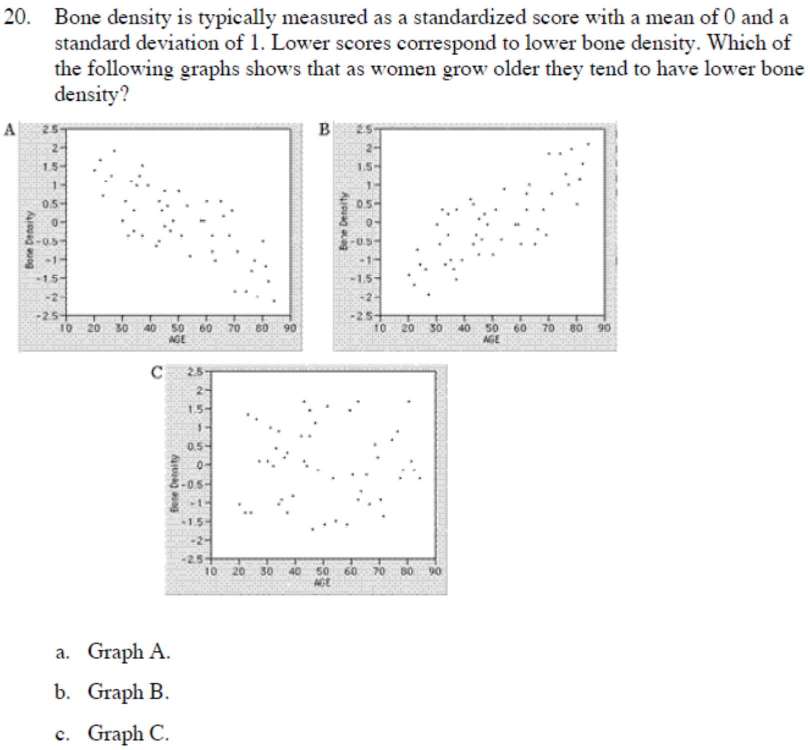
    a. A large *p*-value.

    b. A small *p*-value.

    c. The magnitude of a *p*-value has no impact on statistical significance.

QUESTION 20

Learning Outcome*: Ability to match a scatterplot to a verbal description of a bivariate relationship.*



20. Bone density is typically measured as a standardized score with a mean of 0 and a standard deviation of 1. Lower scores correspond to lower bone density. Which of the following graphs shows that as women grow older they tend to have lower bone density?

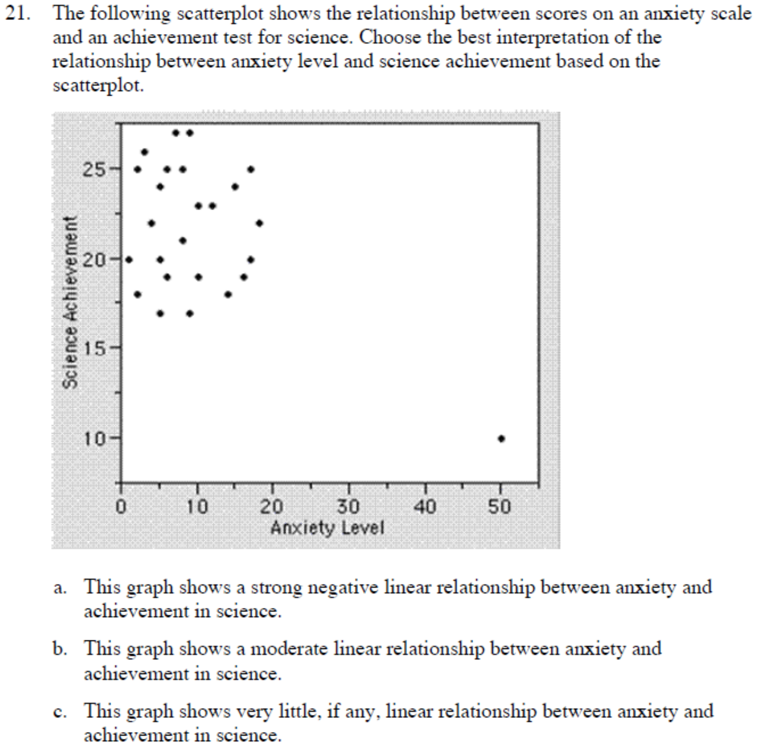
a. Graph A.

b. Graph B.

c. Graph C.

QUESTION 21

Learning Outcome: *Ability to correctly describe a bivariate relationship shown in a scatterplot when there is an outlier (influential point).*



21. The following scatterplot shows the relationship between scores on an anxiety scale and an achievement test for science. Choose the best interpretation of the relationship between anxiety level and science achievement based on the scatterplot.

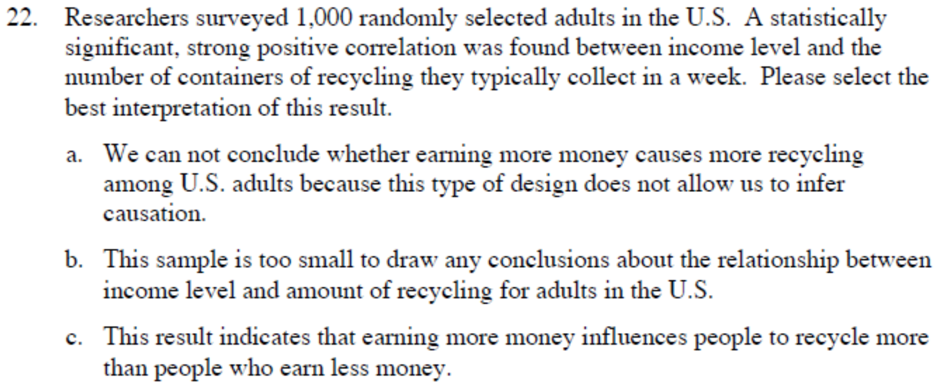
a. This graph shows a strong negative linear relationship between anxiety and achievement in science.

b. This graph shows a moderate linear relationship between anxiety and achievement in science.

c. This graph shows very little, if any, linear relationship between anxiety and achievement in science.

QUESTION 22

Learning Outcome: *Understanding that correlation does not imply causation.*



22. Researchers surveyed 1,000 randomly selected adults in the U.S. A statistically significant, strong positive correlation was found between income level and the number of containers of recycling they typically collect in a week. Please select the best interpretation of this result.

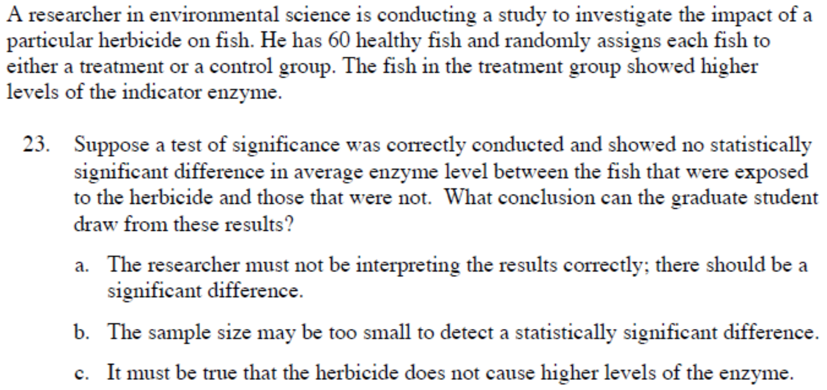
    a. We can not conclude whether earning more money causes more recycling among U.S. adults because this type of design does not allow us to infer causation.

    b. This sample is too small to draw any conclusions about the relationship between income level and amount of recycling for adults in the U.S.

    c. This result indicates that earning more money influences people to recycle more than people who earn less money.

QUESTION 23

Learning Outcome: *Understanding that no statistical significance does not guarantee that there is no effect.*



A researcher in environmental science is conducting a study to investigate the impact of a particular herbicide on fish. He has 60 healthy fish and randomly assigns each fish to either a treatment or a control group. The fish in the treatment group showed higher levels of the indicator enzyme.

23. Suppose a test of significance was correctly conducted and showed no statistically significant difference in average enzyme level between the fish that were exposed to the herbicide and those that were not. What conclusion can the graduate student draw from these results?

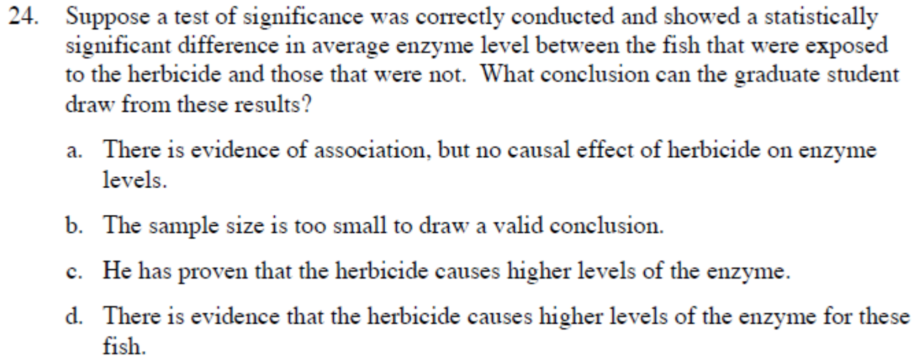
    a. The researcher must not be interpreting the results correctly; there should be a significant difference.

    b. The sample size may be too small to detect a statistically significant difference.

    c. It must be true that the herbicide does not cause higher levels of the enzyme.

QUESTION 24

Learning Outcome: *Understanding that an experimental design with random assignment supports causal inference.*



24. Suppose a test of significance was correctly conducted and showed a statistically significant difference in average enzyme level between the fish that were exposed to the herbicide and those that were not. What conclusion can the graduate student draw from these results?

    a. There is evidence of association, but no causal effect of herbicide on enzyme levels.

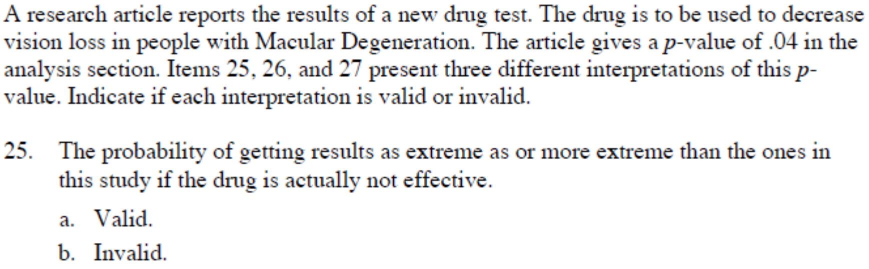
    b. The sample size is too small to draw a valid conclusion.

    c. He has proven that the herbicide causes higher levels of the enzyme.

    d. There is evidence that the herbicide causes higher levels of the enzyme for these fish.

QUESTION 25

Learning Outcome: *Ability to recognize a correct interpretation of a p-value*

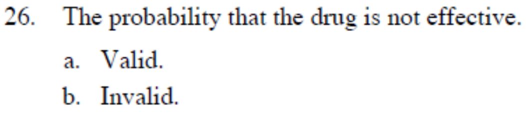


A research article reports the results of a new drug test. The drug is to be used to decrease vision loss in people with Macular Degeneration. The article gives a *p*-value of .04 in the analysis section. Items 25, 26, and 27 present three different interpretations of this *p*-value. Indicate if each interpretation is valid or invalid.

25. The probability of getting results as extreme as or more extreme than the ones in this study if the drug is actually not effective.
    a. Valid.
    b. Invalid.

QUESTION 26

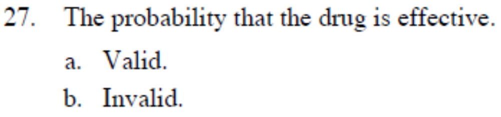
Learning Outcome: *Ability to recognize an incorrect interpretation of a p-value (probability that a treatment is not effective).*



26. The probability that the drug is not effective.
    a. Valid.
    b. Invalid.

QUESTION 27

Learning Outcome: *Ability to recognize an incorrect interpretation of a p-value (prob. treatment as effective).*

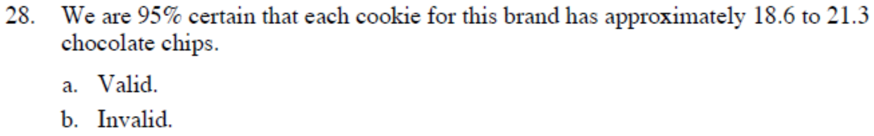


27. The probability that the drug is effective.
    a. Valid.
    b. Invalid.

QUESTION 28

Learning Outcome: *Ability to detect a misinterpretation of a confidence level (the percentage of sample data between confidence limits).*

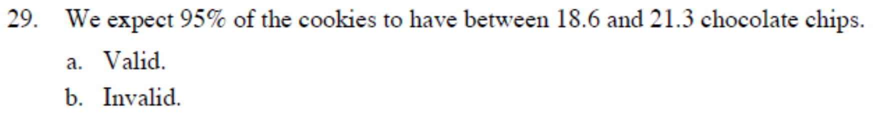
A high school statistics class wants to estimate the average number of chocolate chips in a generic brand of chocolate chips cookies. They collect a random sample of cookies, count the chips in each cookie, and calculate a 95% confidence interval for the average number of chips in each cookie (18.6 to 21.3). Items 28, 29, 30, and 31 present four different interpretations of these results. Indicate if each interpretation is valid or invalid.



28. We are 95% certain that each cookie for this brand has approximately 18.6 to 21.3 chocolate chips.
    a. Valid.
    b. Invalid.

QUESTION 29

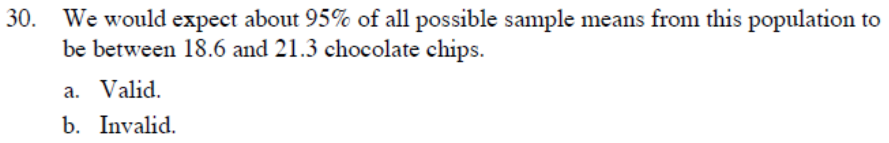
Learning Outcome: *Ability to detect a misinterpretation of a confidence level (percentage of population data values between confidence limits).*



29. We expect 95% of the cookies to have between 18.6 and 21.3 chocolate chips.
    a. Valid.
    b. Invalid.

QUESTION 30

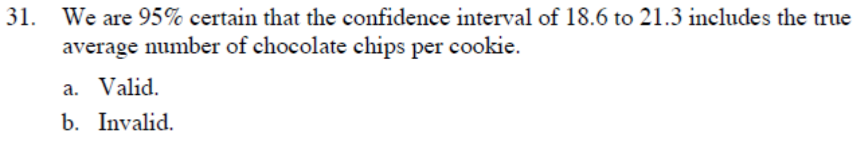
Learning Outcome: *Ability to detect a misinterpretation of a confidence level (percentage of all possible sample means between confidence limits).*



30. We would expect about 95% of all possible sample means from this population to be between 18.6 and 21.3 chocolate chips.
    a. Valid.
    b. Invalid.

QUESTION 31

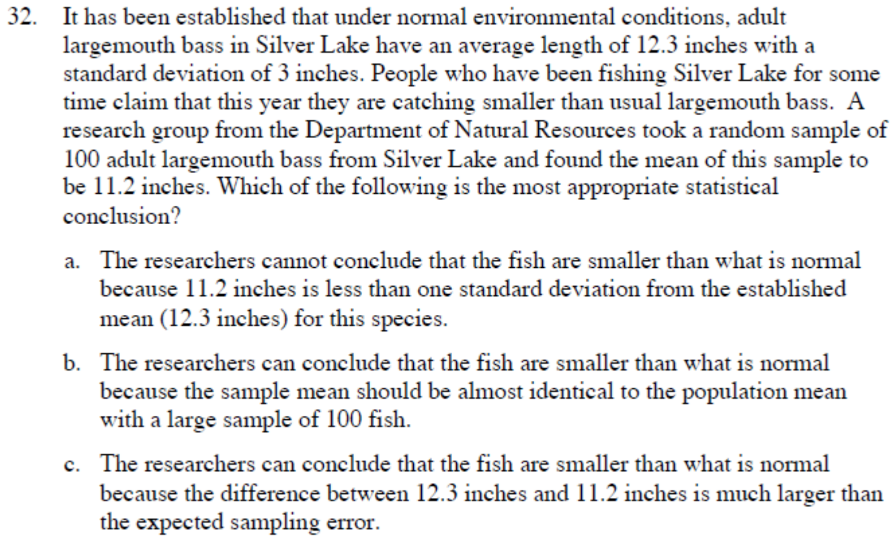
Learning Outcome: *Ability to correctly interpret a confidence interval.*



31. We are 95% certain that the confidence interval of 18.6 to 21.3 includes the true average number of chocolate chips per cookie.
    a. Valid.
    b. Invalid.

QUESTION 32

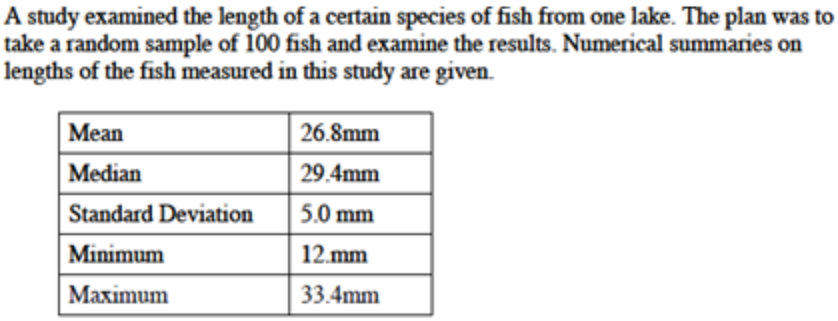
Learning Outcome: *Understanding of how sampling error is used to make an informal inference about a sample mean.*



32. It has been established that under normal environmental conditions, adult largemouth bass in Silver Lake have an average length of 12.3 inches with a standard deviation of 3 inches. People who have been fishing Silver Lake for some time claim that this year they are catching smaller than usual largemouth bass. A research group from the Department of Natural Resources took a random sample of 100 adult largemouth bass from Silver Lake and found the mean of this sample to be 11.2 inches. Which of the following is the most appropriate statistical conclusion?
    a. The researchers cannot conclude that the fish are smaller than what is normal because 11.2 inches is less than one standard deviation from the established mean (12.3 inches) for this species.
    b. The researchers can conclude that the fish are smaller than what is normal because the sample mean should be almost identical to the population mean with a large sample of 100 fish.
    c. The researchers can conclude that the fish are smaller than what is normal because the difference between 12.3 inches and 11.2 inches is much larger than the expected sampling error.

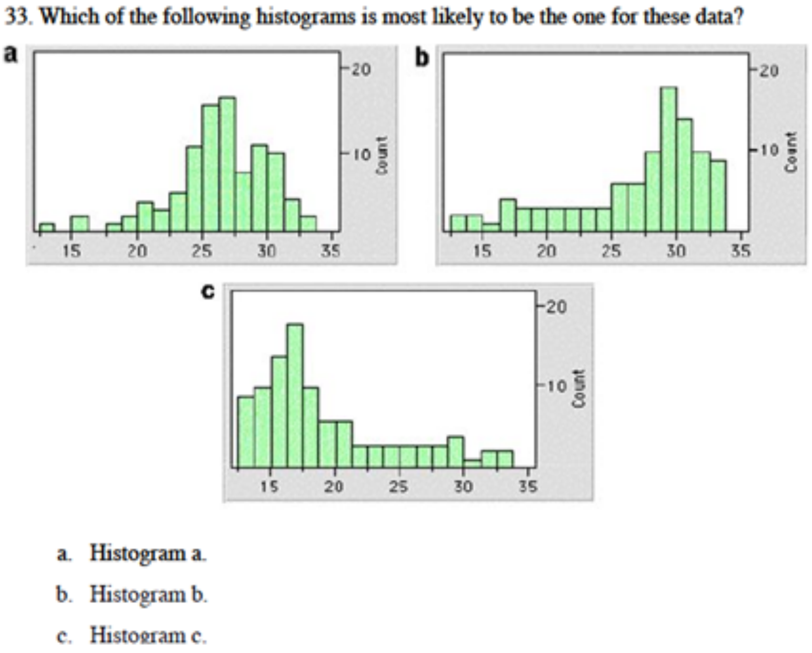
QUESTION 33

Learning Outcome: *Understanding that a distribution with the median larger than the mean is most likely skewed to the left*



A study examined the length of a certain species of fish from one lake. The plan was to take a random sample of 100 fish and examine the results. Numerical summaries on lengths of the fish measured in this study are given.

Mean	26.8mm
Median	29.4mm
Standard Deviation	5.0 mm
Minimum	12.mm
Maximum	33.4mm



33. Which of the following histograms is most likely to be the one for these data?

a. Histogram a.

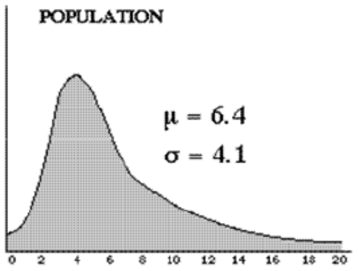
b. Histogram b.

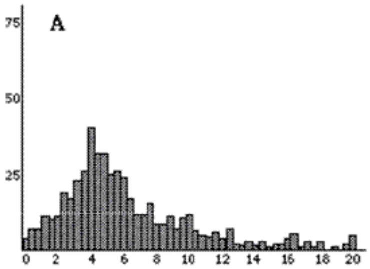
c. Histogram c.

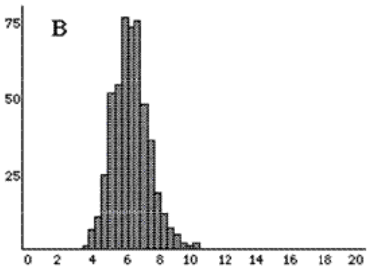
QUESTION 34

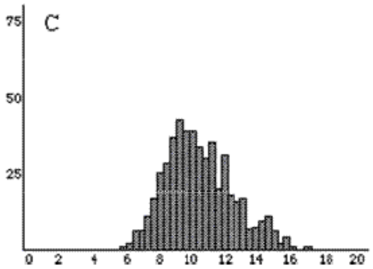
Learning Outcome: *Understanding of the law of large numbers for a large sample by selecting an appropriate sample from a population given the sample size.*

Four graphs are presented below. The graph at the top is a distribution for a population of test scores. The mean score is 6.4 and the standard deviation is 4.1.









34. Which graph (A, B, or C) do you think represents a single random sample of 500 values from this population?

a. Graph A

b. Graph B

c. Graph C

QUESTION 35

Learning Outcome: *Ability to select an appropriate sampling distribution for a population and sample size.*

35. Which graph (A, B, or C) do you think represents a distribution of 500 sample means from random samples each of size 9?

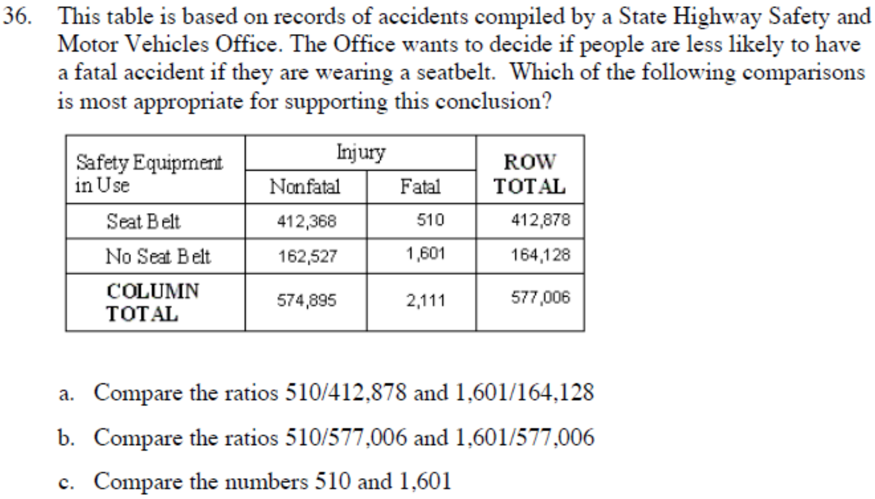
a. Graph A

b. Graph B

c. Graph C

QUESTION 36

Learning Outcome: *Understanding of how to calculate appropriate ratios to find conditional probabilities using a table of data.*



36. This table is based on records of accidents compiled by a State Highway Safety and Motor Vehicles Office. The Office wants to decide if people are less likely to have a fatal accident if they are wearing a seatbelt. Which of the following comparisons is most appropriate for supporting this conclusion?

Safety Equipment in Use	Injury		**ROW TOTAL**
	Nonfatal	Fatal	
Seat Belt	412,368	510	412,878
No Seat Belt	162,527	1,601	164,128
COLUMN TOTAL	574,895	2,111	577,006

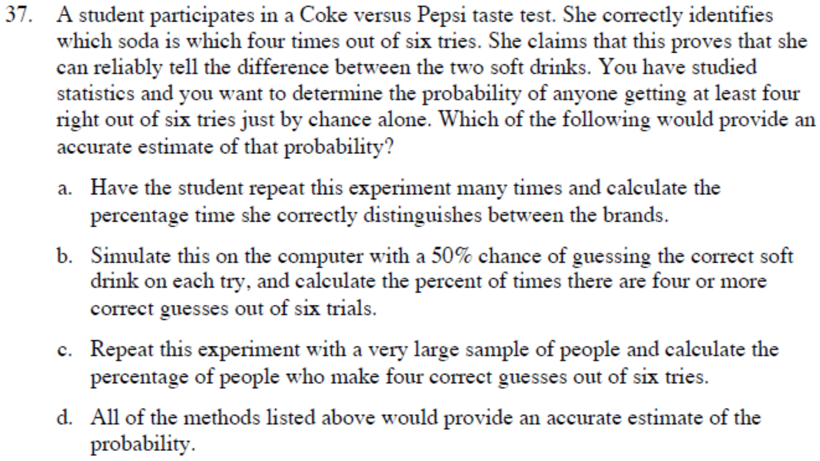
a. Compare the ratios 510/412,878 and 1,601/164,128

b. Compare the ratios 510/577,006 and 1,601/577,006

c. Compare the numbers 510 and 1,601

QUESTION 37

Learning Outcome: *Understanding of how to simulate data to find the probability of an observed value.*



37. A student participates in a Coke versus Pepsi taste test. She correctly identifies which soda is which four times out of six tries. She claims that this proves that she can reliably tell the difference between the two soft drinks. You have studied statistics and you want to determine the probability of anyone getting at least four right out of six tries just by chance alone. Which of the following would provide an accurate estimate of that probability?

a. Have the student repeat this experiment many times and calculate the percentage time she correctly distinguishes between the brands.

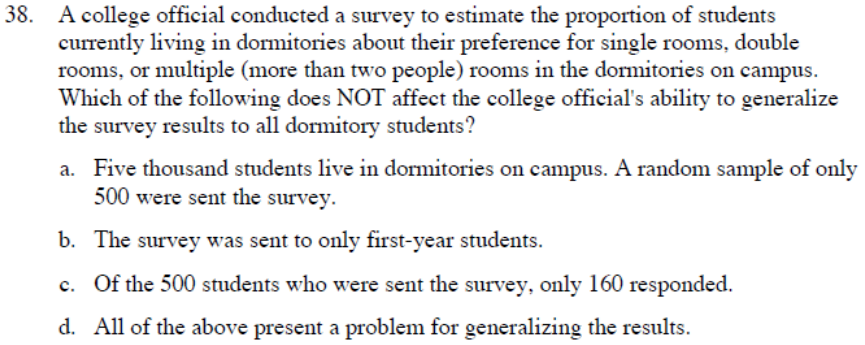
b. Simulate this on the computer with a 50% chance of guessing the correct soft drink on each try, and calculate the percent of times there are four or more correct guesses out of six trials.

c. Repeat this experiment with a very large sample of people and calculate the percentage of people who make four correct guesses out of six tries.

d. All of the methods listed above would provide an accurate estimate of the probability.

QUESTION 38

Learning Outcome: *Understanding of the factors that allow a sample of data to be generalized to the population.*



38. A college official conducted a survey to estimate the proportion of students currently living in dormitories about their preference for single rooms, double rooms, or multiple (more than two people) rooms in the dormitories on campus. Which of the following does NOT affect the college official's ability to generalize the survey results to all dormitory students?

    a. Five thousand students live in dormitories on campus. A random sample of only 500 were sent the survey.

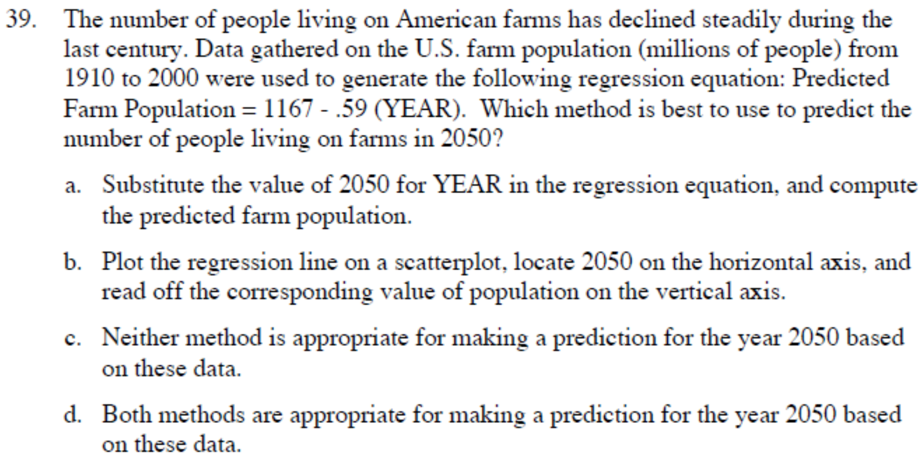
    b. The survey was sent to only first-year students.

    c. Of the 500 students who were sent the survey, only 160 responded.

    d. All of the above present a problem for generalizing the results.

QUESTION 39

Learning Outcome*: Understanding of when it is not wise to extrapolate using a regression model.*



39. The number of people living on American farms has declined steadily during the last century. Data gathered on the U.S. farm population (millions of people) from 1910 to 2000 were used to generate the following regression equation: Predicted Farm Population = 1167 - .59 (YEAR). Which method is best to use to predict the number of people living on farms in 2050?

    a. Substitute the value of 2050 for YEAR in the regression equation, and compute the predicted farm population.

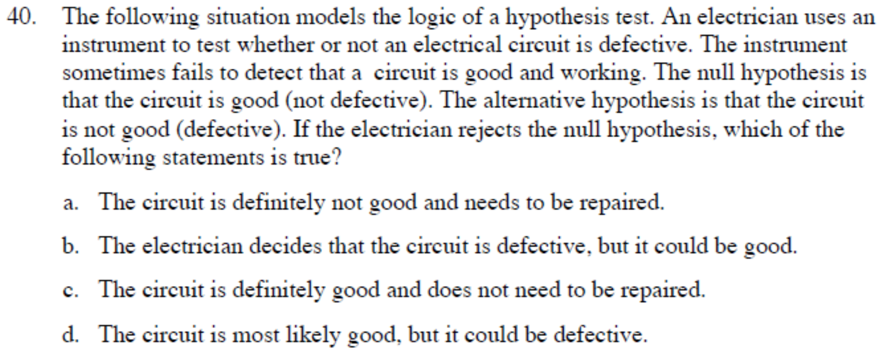
    b. Plot the regression line on a scatterplot, locate 2050 on the horizontal axis, and read off the corresponding value of population on the vertical axis.

    c. Neither method is appropriate for making a prediction for the year 2050 based on these data.

    d. Both methods are appropriate for making a prediction for the year 2050 based on these data.

QUESTION 40

Learning Outcome: *Understanding of the logic of a significance test when the null hypothesis is rejected.*



40. The following situation models the logic of a hypothesis test. An electrician uses an instrument to test whether or not an electrical circuit is defective. The instrument sometimes fails to detect that a circuit is good and working. The null hypothesis is that the circuit is good (not defective). The alternative hypothesis is that the circuit is not good (defective). If the electrician rejects the null hypothesis, which of the following statements is true?

    a. The circuit is definitely not good and needs to be repaired.

    b. The electrician decides that the circuit is defective, but it could be good.

    c. The circuit is definitely good and does not need to be repaired.

    d. The circuit is most likely good, but it could be defective.

CAOS 4 ANSWER KEY

1. D	11. B	21. C	31. A
2. B	12. A	22. A	32. C
3. C	13 B	23. B	33. B
4. A	14. A	24. D	34 A
5. D	15. B	25. A	35. B
6. C	16. C	26. B	36. A
7. D	17 A	27. B	37. B
8. A	18. A	28. B	38. A
9. D	19. B	29. B	39. C
10. C	20. A	30. B	40. B

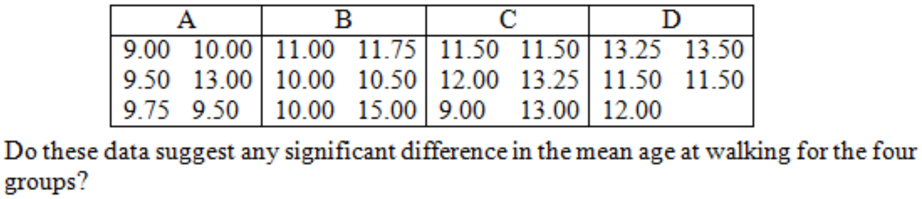
Appendix F – 24 problems – Studies 5 & 6

1) One independent variable + Continuous dependent variable *[One-way ANOVA]*

Health

Problem #5

Zelazo et al. (1972) report on an experiment to determine the effect of special walking exercises on the age at which children begin to walk. Twenty-three infants were randomly divided into four groups. Those infants in groups A and B received various exercises, whereas those in groups C and D did not. The following data are the ages (in months) at which each of the 23 children first walked.

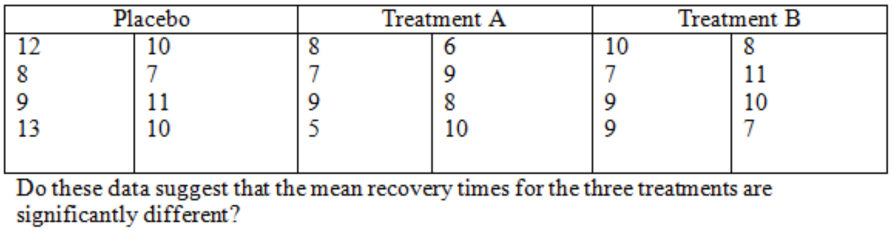


A		B		C		D	
9.00	10.00	11.00	11.75	11.50	11.50	13.25	13.50
9.50	13.00	10.00	10.50	12.00	13.25	11.50	11.50
9.75	9.50	10.00	15.00	9.00	13.00	12.00	

Do these data suggest any significant difference in the mean age at walking for the four groups?

Problem #10

A clinic randomly assigns 24 patients suffering from blisters to receive one of three treatments, one of which is a placebo. The number of days for the blisters to completely heal are as follows:



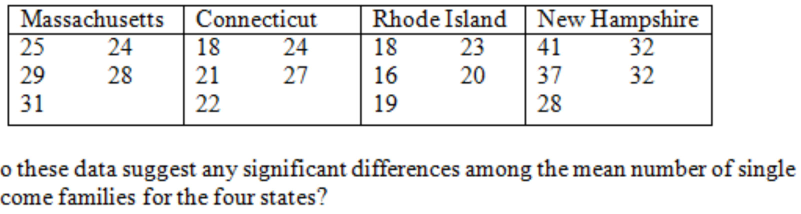
Placebo		Treatment A		Treatment B	
12	10	8	6	10	8
8	7	7	9	7	11
9	11	9	8	9	10
13	10	5	10	9	7

Do these data suggest that the mean recovery times for the three treatments are significantly different?

Demographics

Problem #21

A random sample of 20 communities in New England is selected. The following figures are the number of single income families for each community by state.

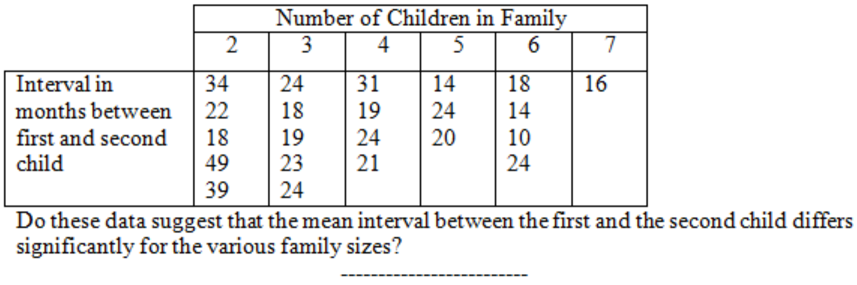


Massachusetts		Connecticut		Rhode Island		New Hampshire	
25	24	18	24	18	23	41	32
29	28	21	27	16	20	37	32
31		22		19		28	

Do these data suggest any significant differences among the mean number of single income families for the four states?

Problem #22

A demographer is interested in the relationship between the birth of a family's first and second child and the eventual family size. She follows 22 families for 20 years, collecting the following data:



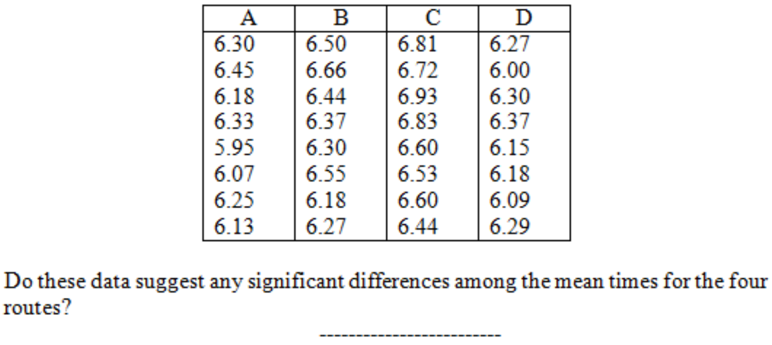
	Number of Children in Family					
	2	3	4	5	6	7
Interval in months between first and second child	34	24	31	14	18	16
	22	18	19	24	14	
	18	19	24	20	10	
	49	23	21		24	
	39	24				

Do these data suggest that the mean interval between the first and the second child differs significantly for the various family sizes?

Products and Services

Problem #20

A bus company plans to begin service between two cities. Four routes, A, B, C, and D, are under consideration. To assess differences in the mean time for the four routes, a bus makes the trip between the cities 32 times, taking each route eight times. The times (in hours) for each trip are as follows:

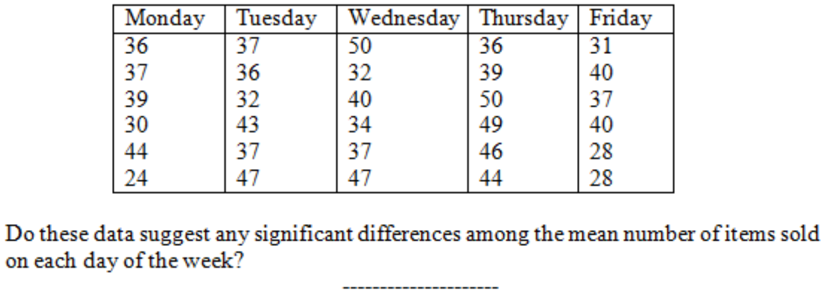


A	B	C	D
6.30	6.50	6.81	6.27
6.45	6.66	6.72	6.00
6.18	6.44	6.93	6.30
6.33	6.37	6.83	6.37
5.95	6.30	6.60	6.15
6.07	6.55	6.53	6.18
6.25	6.18	6.60	6.09
6.13	6.27	6.44	6.29

Do these data suggest any significant differences among the mean times for the four routes?

Problem #23

A company records the number of items sold by their salespersons on a sample of six Mondays, six Tuesdays, and so on. The results are as follows:



Monday	Tuesday	Wednesday	Thursday	Friday
36	37	50	36	31
37	36	32	39	40
39	32	40	50	37
30	43	34	49	40
44	37	37	46	28
24	47	47	44	28

Do these data suggest any significant differences among the mean number of items sold on each day of the week?

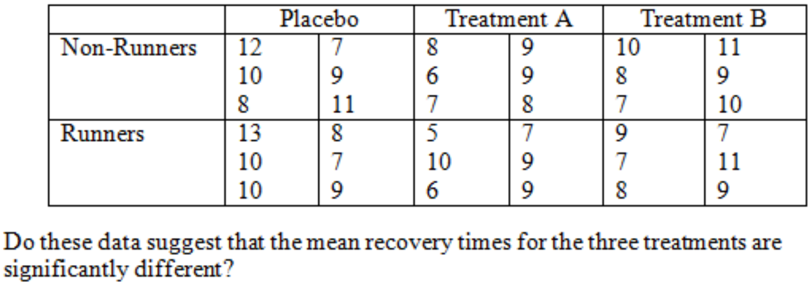
----

2) Two independent variables + Continuous dependent variable *[Two-Way ANOVA]*

Health

Problem #16

A clinic randomly assigns *42* patients suffering from blisters to receive one of three treatments, one of which is a placebo. The patients are both *runners and non-runners*. The number of days for the blisters to completely heal are as follows:

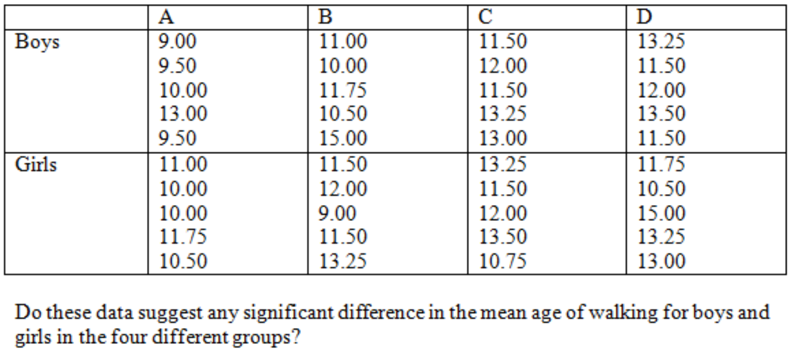


	Placebo		Treatment A		Treatment B	
Non-Runners	12	7	8	9	10	11
	10	9	6	9	8	9
	8	11	7	8	7	10
Runners	13	8	5	7	9	7
	10	7	10	9	7	11
	10	9	6	9	8	9

Do these data suggest that the mean recovery times for the three treatments are significantly different?

Problem #19

Zelazo et al. (1972) report on an experiment to determine the effect of special walking exercises on the age at which boys and girls begin to walk. Forty infants were randomly divided into four groups. Those infants in groups A and B received various exercises, whereas those in groups C and D did not. The following data are the ages (in months) at which each of the 40 children first walked.



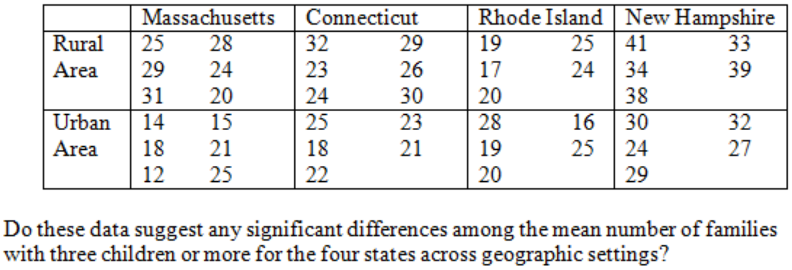
	A	B	C	D
Boys	9.00	11.00	11.50	13.25
	9.50	10.00	12.00	11.50
	10.00	11.75	11.50	12.00
	13.00	10.50	13.25	13.50
	9.50	15.00	13.00	11.50
Girls	11.00	11.50	13.25	11.75
	10.00	12.00	11.50	10.50
	10.00	9.00	12.00	15.00
	11.75	11.50	13.50	13.25
	10.50	13.25	10.75	13.00

Do these data suggest any significant difference in the mean age of walking for boys and girls in the four different groups?

Demographics

Problem #4

A random sample of 43 communities in New England is selected. The following figures are the number of families with three children or more for each community by state and geographic setting.

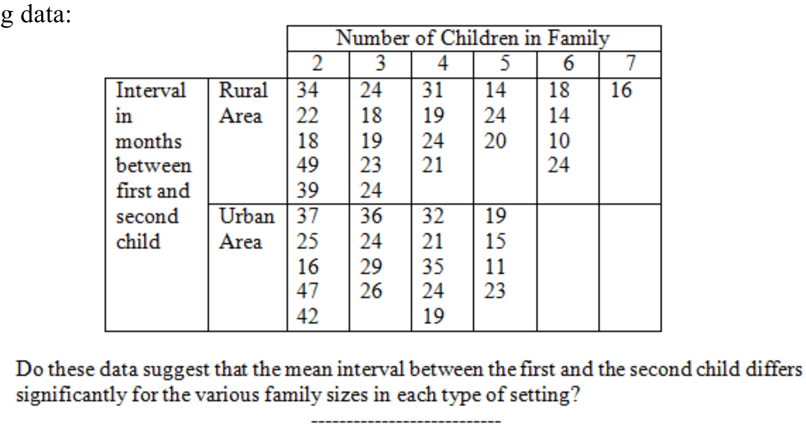


	Massachusetts		Connecticut		Rhode Island		New Hampshire	
Rural Area	25	28	32	29	19	25	41	33
	29	24	23	26	17	24	34	39
	31	20	24	30	20		38	
Urban Area	14	15	25	23	28	16	30	32
	18	21	18	21	19	25	24	27
	12	25	22		20		29	

Do these data suggest any significant differences among the mean number of families with three children or more for the four states across geographic settings?

Problem #8

A demographer is interested in the relationship between the birth of a family's first and second child and the eventual family size, both in rural and in urban settings. She follows 40 families for 20 years, collecting the following data:

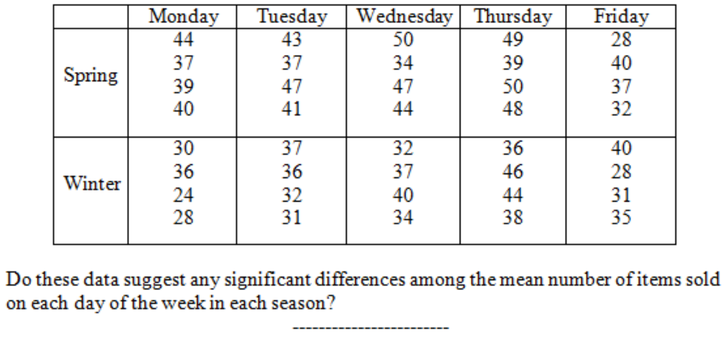


		Number of Children in Family					
		2	3	4	5	6	7
Interval in months between first and second child	Rural Area	34	24	31	14	18	16
		22	18	19	24	14	
		18	19	24	20	10	
		49	23	21		24	
		39	24				
	Urban Area	37	36	32	19		
		25	24	21	15		
		16	29	35	11		
		47	26	24	23		
		42		19			

Do these data suggest that the mean interval between the first and the second child differs significantly for the various family sizes in each type of setting?

Products and Services

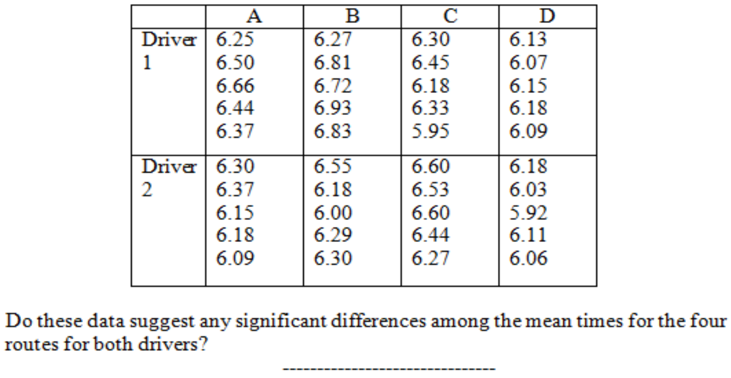
Problem #1
A company records the number of items sold by their salespersons on a sample of six Mondays, six Tuesdays, and so on. Of the eight sales figures for each day, four referred to days in spring and four referred to days in winter. The results are as follows:



	Monday	Tuesday	Wednesday	Thursday	Friday
Spring	44 37 39 40	43 37 47 41	50 34 47 44	49 39 50 48	28 40 37 32
Winter	30 36 24 28	37 36 32 31	32 37 40 34	36 46 44 38	40 28 31 35

Do these data suggest any significant differences among the mean number of items sold on each day of the week in each season?

Problem #17
A bus company plans to begin service between two cities. Four routes, A, B, C, and D, are under consideration. To assess differences in the mean time for the four routes, two bus drivers make the trip between the cities 20 times, each driver taking each route five times. The times (in hours) for each trip are as follows:



	A	B	C	D
Driver 1	6.25 6.50 6.66 6.44 6.37	6.27 6.81 6.72 6.93 6.83	6.30 6.45 6.18 6.33 5.95	6.13 6.07 6.15 6.18 6.09
Driver 2	6.30 6.37 6.15 6.18 6.09	6.55 6.18 6.00 6.29 6.30	6.60 6.53 6.60 6.44 6.27	6.18 6.03 5.92 6.11 6.06

Do these data suggest any significant differences among the mean times for the four routes for both drivers?

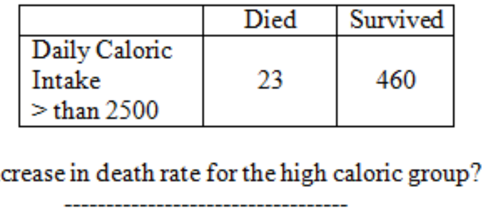
----

3) One independent variable + Categorical dependent variable *[Test of Fit, Chi-Square]*

Health

Problem #11

Part of the Framingham heart study (Gordon et al., 1981) focused on the number of deaths from heart attack or heart disease among men aged 45-64. Data show that 7% of those who had a daily caloric intake below 2500 died during the study. The following table gives the number of deaths from heart attack or heart disease among men aged 45-64 who consumed more than 2500 calories daily:

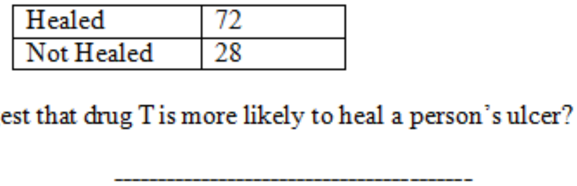


	Died	Survived
Daily Caloric Intake > than 2500	23	460

Do these data suggest an increase in death rate for the high caloric group?

Problem #24

A new antiulcer drug, T, is being promoted as the more efficient one on the market. The efficiency of the current leading drug, Z, is estimated at 64% healing within a month. A group of 200 persons suffering with duodenal ulcers are given drug T. The number of cases in which the ulcers healed within a month were as follows:



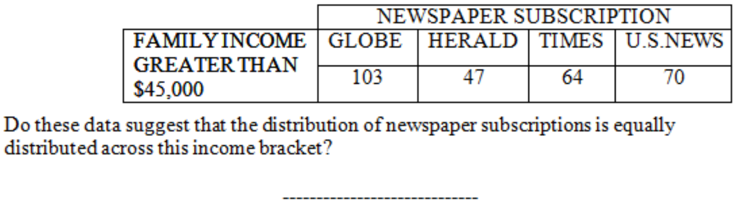
Healed	72
Not Healed	28

Do these data suggest that drug T is more likely to heal a person's ulcer?

Demographics

Problem #13

A marketing organization selected a random sample of adults in a city. Each respondent is classified by daily newspapers read. (The few persons regularly receiving no or two more daily newspapers were omitted.) Their client wanted to know if the distribution of newspaper readership is equally distributed among the high-end consumers (family income greater than $45,000) as they are getting ready for a print ad campaign:

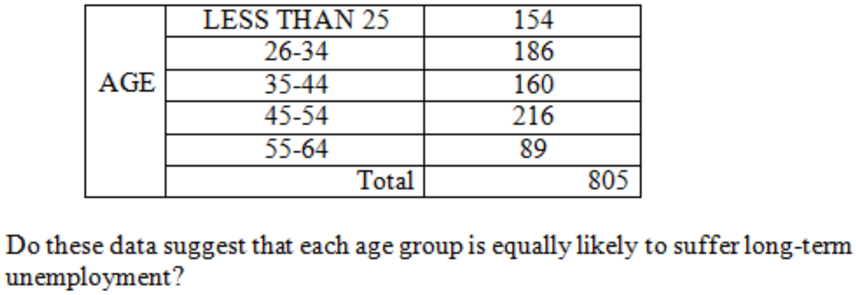


	NEWSPAPER SUBSCRIPTION			
FAMILY INCOME GREATER THAN $45,000	GLOBE	HERALD	TIMES	U.S.NEWS
	103	47	64	70

Do these data suggest that the distribution of newspaper subscriptions is equally distributed across this income bracket?

Problem #18

Each of a sample of 805 adult males who have been unemployed for at least 6 months is classified on the basis of age. The results are summarized as follows:



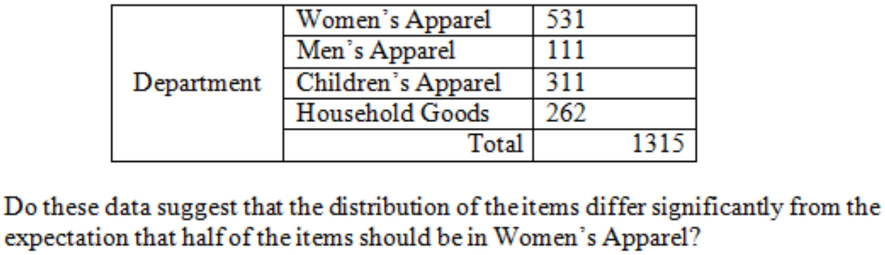
AGE	LESS THAN 25	154
	26-34	186
	35-44	160
	45-54	216
	55-64	89
	Total	805

Do these data suggest that each age group is equally likely to suffer long-term unemployment?

Products and Services

Problem #7

A random sample of 1315 items in the stockroom of a clothing store is classified by department. The results are as follows:

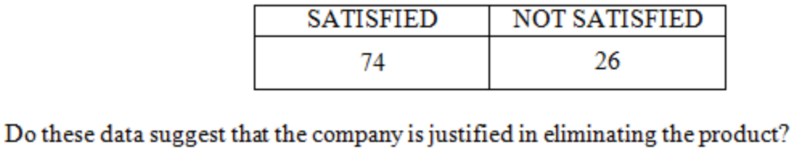


Department	Women's Apparel	531
	Men's Apparel	111
	Children's Apparel	311
	Household Goods	262
	Total	1315

Do these data suggest that the distribution of the items differ significantly from the expectation that half of the items should be in Women's Apparel?

Problem #12

A random sample of 120 customers who had purchased a new product at a department store was asked whether they were satisfied with the product. The company will decide to keep the product on the market only if satisfaction reaches 80%. The results are below:



SATISFIED	NOT SATISFIED
74	26

Do these data suggest that the company is justified in eliminating the product?

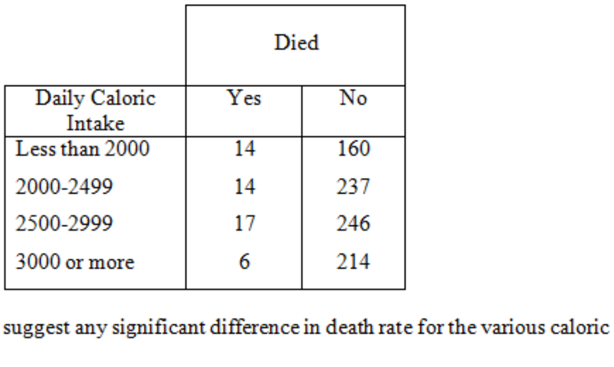
----

4) <u>Two independent variables + Categorical dependent variable</u>
<u>*[Test of Independence, Chi-Square]*</u>

Health

Problem #6

Gordon et al. (1981) report on some of the outcomes of the Framingham heart study. The following table gives the number of deaths from heart attack or heart disease among men aged 45-64 broken down by daily calorie intake:

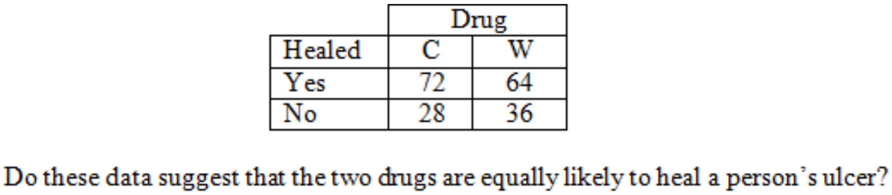


	Died	
Daily Caloric Intake	Yes	No
Less than 2000	14	160
2000-2499	14	237
2500-2999	17	246
3000 or more	6	214

Do these data suggest any significant difference in death rate for the various caloric groups?

Problem #9

Two antiulcer drugs, C and W, are to be compared. A group of 200 persons suffering with stomach ulcers are randomly divided into two groups of 100 each. The members of group I are given drug C and members of group II, drug W. The number of cases in which the ulcers healed within a month were as follows:



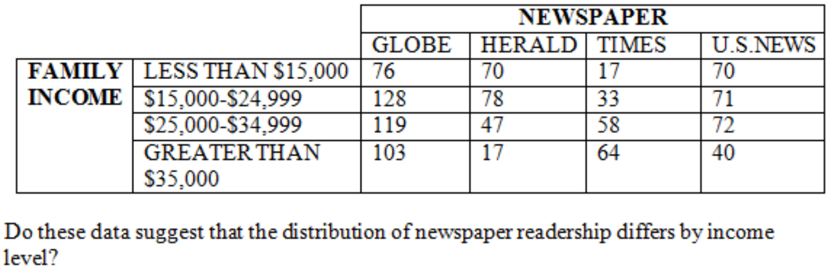
	Drug	
Healed	C	W
Yes	72	64
No	28	36

Do these data suggest that the two drugs are equally likely to heal a person's ulcer?

Demographics

Problem #3

A survey organization selected a random sample of adults in a city. Each respondent is classified by annual income of the family and by daily newspapers read. (The few persons regularly receiving no or two more daily newspapers were omitted.) The results are summarized as follows:

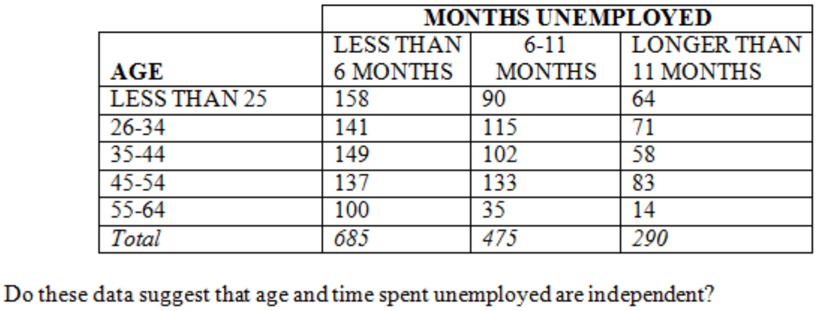


		NEWSPAPER			
		GLOBE	HERALD	TIMES	U.S.NEWS
FAMILY INCOME	LESS THAN $15,000	76	70	17	70
	$15,000-$24,999	128	78	33	71
	$25,000-$34,999	119	47	58	72
	GREATER THAN $35,000	103	17	64	40

Do these data suggest that the distribution of newspaper readership differs by income level?

Problem #15

Each of a sample of 1450 currently unemployed adult males is classified on the basis of age and number of months unemployed. The results are summarized as follows:



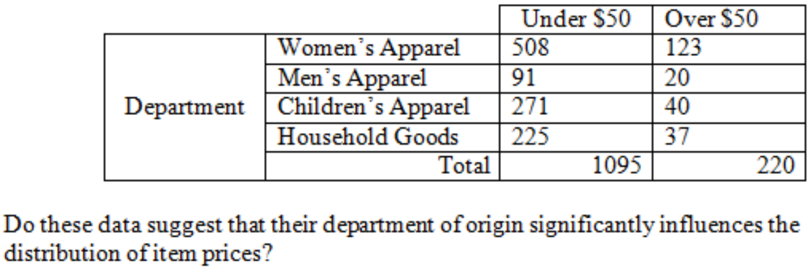
	MONTHS UNEMPLOYED		
AGE	LESS THAN 6 MONTHS	6-11 MONTHS	LONGER THAN 11 MONTHS
LESS THAN 25	158	90	64
26-34	141	115	71
35-44	149	102	58
45-54	137	133	83
55-64	100	35	14
Total	*685*	*475*	*290*

Do these data suggest that age and time spent unemployed are independent?

Products & Services

Problem #2

A random sample of 1315 items in the stockroom of a clothing store is classified by price ('Under $50' or 'Over $50') and by department. The results are as follows:

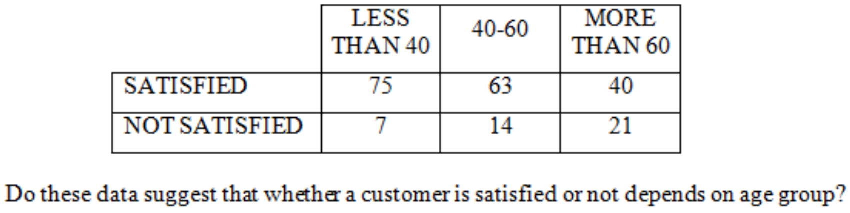


		Under $50	Over $50
Department	Women's Apparel	508	123
	Men's Apparel	91	20
	Children's Apparel	271	40
	Household Goods	225	37
	Total	1095	220

Do these data suggest that their department of origin significantly influences the distribution of item prices?

Problem #14

A random sample of 220 customers who had made purchases at a department store were asked whether they were satisfied with the service. The results were broken down by age group.



	LESS THAN 40	40-60	MORE THAN 60
SATISFIED	75	63	40
NOT SATISFIED	7	14	21

Do these data suggest that whether a customer is satisfied or not depends on age group?
