THE ROLES OF EXPERIENCE, GENDER, AND INDIVIDUAL DIFFERENCES IN STATISTICAL REASONING

NADIA MARTIN
University of Waterloo
nadia.martin@uwaterloo.ca

JEFFREY HUGHES
University of Waterloo
j4hughes@uwaterloo.ca

JONATHAN FUGELSANG
University of Waterloo
jafugels@uwaterloo.ca

ABSTRACT

We examine the joint effects of gender and experience on statistical reasoning. Participants with various levels of experience in statistics completed the Statistical Reasoning Assessment (Garfield, 2003), along with individual difference measures assessing cognitive ability and thinking dispositions. Although the performance of both genders improved with experience, the gender gap persisted, with males outperforming females across all experience levels. A confirmatory structural equation model assessing the degree to which cognitive ability, thinking dispositions, and gender account for statistical reasoning performance supported the idea that differences in statistical reasoning are not uniquely a matter of cognitive ability. Rather, gender was found to influence statistical reasoning directly, as well as indirectly through its influence on thinking dispositions.

Keywords: *Statistics education research; Cognitive ability; Thinking dispositions*

1. INTRODUCTION

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write!
--S.S. Wilks (1951)

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 (Ben-Zvi & Garfield, 2008; Rumsey, 2002; Wallman, 1993). The new door to knowledge is data (Lohr, 2009) and statistical competence holds the key to that door. Statistical competence—numeracy paired with critical thinking—allows for proper evaluation of data to guide decision- and policy-making. Today, a low level of numeracy is detrimental to informed decision-making

Statistics Education Research Journal, 16(2), XX-XX, http://iase-web.org/Publications.php?p=SERJ

(e.g., choosing between two medical treatments: Couper & Singer, 2009; McHugh & Behar, 2009; Reyna, Nelson, Han, & Dieckmann, 2009), and to employability, with outcomes potentially worse for women than men (Parsons & Bynner, 1997, 2005). The current study focuses on the role of gender in the development of statistical competence, with a focus on statistical reasoning—the ability to interpret statistical information and to make decisions based on it (Garfield, 2002; Garfield & Gal, 1999). To get a clearer picture of the development of statistical reasoning, we also take into consideration the role of experience and individual differences in cognitive ability and thinking dispositions.

1.1. STATISTICAL REASONING

The new faces of work and access to information have already influenced the structure of education in the field of statistics. In the early 1990s, at a time when dissatisfaction was growing with introductory statistics classes and when technology was becoming more prevalent, a statistics education reform in the United States was being discussed. The release of the "Cobb Report" (1992) also highlighted the need to acknowledge the changing face of technology, the need to emphasize thinking over procedures, as well as the need to recognize the variety of students accessing statistics classes. This launched further discussion and, a decade later, the Guidelines for Assessment and Instruction in Statistics Education (GAISE) (American Statistical Association, 2005) was released. In particular, the GAISE report advocates conceptual understanding over mere procedural knowledge. Some of the learning goals in this new era of statistical education include understanding the purpose and logic of statistical investigations, learning statistical skills such as organizing data and constructing tables, and developing useful statistical dispositions, such as demonstrating critical reasoning when assessing evidence.

All those learning goals can be related to the concepts of statistical literacy, reasoning, and/or thinking. Rumsey (2002) describes statistical literacy as the basic knowledge supporting the ability to consume daily statistical information. In contrast, Garfield (2002) defines statistical reasoning as interpreting statistical information and making decisions based on it. Finally, Chance (2002) describes statistical thinking as mental habits and questioning tendencies, encompassing the ability to see the statistical process as a whole and to move beyond the textbook. However, all three authors highlighted the fact that each aspect has been defined and used inconsistently in the literature, with much overlap in their definitions. Other fields, such as psychology, also do not always make a strong distinction between reasoning and thinking, with the two terms often being used interchangeably. Whereas reasoning relates to formal thinking practices such as drawing inferences and deductions (Evans, 2002), thinking is generally a broader umbrella term that refers to going beyond the information given (Bruner, 1957), thus encompassing the act of reasoning. Noting how the distinction between statistical literacy, reasoning, and thinking can be subtle, with many overlapping features, delMas (2002) argued that they are potentially easier to identify at the assessment level than in their definitions. This is due to the fact that the actions required in the assessment task will reveal which aspect is under evaluation. For instance, whereas asking for a simple calculation, identification, or description could be sufficient to test *literacy*, testing *reasoning* could require explaining how or why a result was obtained (delMas, Garfield, & Ooms, 2005). In contrast, whereas both literacy and reasoning may be assessed with neutral content, testing statistical

thinking should be done with the use of a context, such as asking a student to evaluate and critique a study design and conclusions.

1.2. ASSESSING STATISTICAL REASONING

With the introduction of the GAISE report came the need to measure the impact of its recommendations on students' learning, an important goal for educators and researchers in the field of statistics. One evaluation tool that was developed specifically to measure statistical reasoning is the Statistical Reasoning Assessment (SRA) (Garfield, 1991, 1998, 2002, 2003; Garfield & Gal, 1999). The SRA 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. Designed to assess a wide range of statistical concepts covered in high school and in introductory statistics classes at the college or university 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, interpretation, or understanding that are not only mistaken (despite often being intuitively plausible), but also resistant to change (Chi & Roscoe, 2002; Fischbein, 1987) and can be quite 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 trust small samples to be representative of the population (Kahneman, Slovic, & Tversky, 1982) and base their decisions on them.

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 for both classroom assessments and research. In this multiple-choice format, the true answers are embedded amongst incorrect answers (foils) whose content is based on erroneous but plausible answers given by actual students in an early round of the instrument's development.

It is the multiple-choice format of the SRA that allows for the easy evaluation of the use of correct reasoning versus misconceptions. For each question, some of the choices represent correct reasoning, whereas other choices represent some prevalent misconceptions. As participants are not instructed to choose only one answer, both types of answers can be detected in each question. Its short length and ease of scoring allows educators to quickly measure development and achievement in the classroom, even for instructors of large classes (Garfield & Chance, 2000). It also provides an efficient tool for researchers, especially for laboratory research where practical considerations such as duration of a session must be taken into account. Although there are other assessment tools, such as the CAOS (delMas, Garfield, Ooms, & Chance, 2007), which aims to assess "students' conceptual understanding of important statistical ideas" with 40 questions, the SRA was judged to represent an appropriate and advantageous choice in our situation, as multiple instruments had to be administered within

a limited time period. In addition, to our knowledge, published research with the CAOS has not directly compared the performance of males and females.

The inclusion of many different areas of understanding within one single research tool breaks from the tradition of much published research, especially in psychology. It is common to read articles focusing on a single aspect of statistical competence such as the law of large numbers (e.g., Fong, Krantz, & 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 necessarily makes for a relatively low internal consistency, its test-retest reliability of 0.70 for the correct reasoning scale and 0.75 for the misconception scale (Garfield, 2003) makes it a good choice for research (Nunnally & Bernstein, 1994).

1.3. VARIABLES AFFECTING STATISTICAL REASONING

Gender Several researchers have used the SRA to assess statistical competence in a wide range of populations, and numerous important findings have emerged. Critically for our 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 also marginally significant for correct reasoning. Also using the SRA—though administering it at the beginning of a semester before any statistics instruction in contrast to Garfield who tested their participants at the end of an introductory course in statistics—Tempelaar, Gijselaers, and Schim van der Loeff (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 ($p < .001$, $d = .27$). Tempelaar et al. also discovered a weak negative correlation between the SRA and effort-based measures (i.e., homework), and a weak positive correlation between the correct reasoning score on the SRA and the final exam. In contrast, Garfield found no correlation between performance on the SRA and course performance. As Tempelaar et al. note, what is especially puzzling is the fact that this gender difference occurs despite similar educational backgrounds for the males and females. Then again, other 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. Given the previous findings, as our first hypothesis, we also expect that males will outperform females on the SRA. However, because both Garfield and Tempelaar et al. limited the range of experience in their research, the question of knowing whether the gender gap is persistent or transient remains unanswered.

Experience 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, different domains of study (e.g., chemistry, psychology, law) tend to emphasize and develop dissimilar reasoning skills (Gray & Mill, 1990; Nisbett, Fong, Lehman, & Cheng, 1987). Fong et al.'s (1986) findings also fail to control for cognitive ability and motivation. 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 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. Quilici and Mayer demonstrated that appropriate training in statistics led participants to sort statistical word problems based on their deep structure rather than based on their surface similarity. 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. For those interested in statistical competence 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 competent in today's society. In contrast to the tasks used in the studies above, the SRA addresses multiple areas of statistical reasoning. Of course, such breadth of knowledge cannot be communicated in a single training session. Thus, our hypothesis is that considerably more training will be necessary to generate a significant improvement. However, given that training studies have not reported gender and that gender-reporting studies have limited the range of experience, we can only speculate on the potential interaction between gender and experience. On the one hand, it is possible that the gender gap will remain despite increased experience. On the other hand, it is possible that the gap will diminish. Empirical data is needed to elucidate this question. We aim to answer this question in the current study by including both males and females with various levels of experience in statistics.

Cognitive Ability and Thinking Dispositions As noted above, studies of statistical reasoning concerned with the role of gender or experience typically have not controlled for cognitive ability or motivation. In reasoning research, motivation can be associated with thinking dispositions, which can be described simply as intellectual inclinations that benefit good, productive thinking (Ritchhart, 2001). Thus, a second goal of the current study was to better understand the role that individual differences in cognitive ability and thinking dispositions can have on statistical reasoning. Such individual differences approaches have 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., Baron, 1985; Ennis, 1987; Stanovich, 1999). Indeed, Stanovich and West's research (1997, 1998a, 1998b, 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. In support of the role of thinking dispositions in statistical reasoning, Hawkins (1997) states

> 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' [sic] 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] (p. 13)

Wild and Pfannkuch (1999) have also discussed the importance of acknowledging the role of thinking dispositions in statistics, recognizing that students will not all equally engage with a problem. In addition to one's baseline interest in a topic, certain personal qualities such as curiosity, propensity to seek deeper meaning and to question conclusions, as well as openness to ideas that challenge one's preconceptions, will all influence the degree to which a person will be willing to engage with and think about a problem. Further, Chance (2002) emphasizes that, even though thinking statistically clearly requires inquisitive "habits of mind" such as being sceptical about the data obtained, those are more likely to emerge and become permanent when instruction makes them explicit often.

In fact, another reason that thinking dispositions are an attractive focus for research is because they are seen as more malleable than cognitive abilities (Baron, 1985; Stanovich, 1999) 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 task can be improved simply by influencing their level of motivation and dedication to the task. Similarly, 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 (2009) gives to thinking dispositions in his influential model of reasoning. 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 we are aware, the appropriateness of this model for statistical reasoning has not been tested directly. Thus, in line with our second goal, we examined the role of thinking dispositions and cognitive abilities in statistical reasoning using a confirmatory structural equation modeling approach. We hypothesized that increases in thinking dispositions

would have a positive impact on statistical reasoning, above and beyond the contributions of cognitive abilities.

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.

2. METHOD

2.1. PARTICIPANTS

Two hundred and one undergraduate (76%) and graduate (24%) students proficient in English voluntarily participated for course credit or monetary remuneration between September 2010 and August 2011. Undergraduates were recruited through the standard psychology department participant pool, and graduate students were recruited through an email sent to the entire graduate students' email list. Participants had varying levels of experience in statistics, as measured by the self-reported number of statistics courses they had previously taken (0: $n = 76$, 1: $n = 46$, 2: $n = 38$, 3: $n = 15$, 4 or more: $n = 24$). Graduate students were included in the study to increase the number of participants having completed at least two courses in statistics and allow the testing of our hypotheses related to experience. All those having taken 2, 3, or 4 courses were grouped under experience level 2+ to create more equal cell sizes. As participants came from a variety of programs and courses, they were also asked to report the level of quantitative knowledge required in their field (questionnaire item #9 from Schield, 2005), reporting this value on a 5-point scale ranging from 'generally non-quantitative' (e.g., child care, music, art, English, philosophy), to 'minimally quantitative' (e.g., business management, education, journalism, health care), to 'moderately quantitative' (e.g., psychology, sociology, market research, forecasting), to 'highly quantitative' (e.g., finance, econometrics, accounting, science, engineering), to 'extremely quantitative' (e.g., mathematics, statistics). Most participants were from 'highly quantitative' (41%) and 'moderately quantitative' (30%) fields, followed by 'extremely quantitative' (13%), 'minimally quantitative' (11%) and 'generally non-quantitative' (5%) fields. Following Frederick (2005), two participants with scores below 10 on the Wonderlic Personnel Test (see Section 2.2) were eliminated from the analysis, reducing the sample to 199 participants (46% males, 54% females; $mean_{age} = 21.57$ years, $SD_{age} = 4.59$ years).

2.2. DESIGN AND MATERIALS

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

Statistical Task The 20-item Statistical Reasoning Assessment (SRA; Garfield, 2003) was used as the main task. 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). Whereas a high score on the CR scale indicates better performance, it is a low score on the MISC scale that is indicative of better performance (i.e., better avoidance of misconceptions). Descriptive statistics are reported in Table 1.

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 The following measures of thinking dispositions and measures of cognitive ability were used (see Table 1 for descriptive statistics): the 18-item Need for Cognition Scale (NC; Cacioppo et al.,1984), the 20-item Preference for Numerical Information Scale (PNI; Viswanathan, 1993), and the 41-item Actively Open-Minded Thinking Scale (AOT; Sà, West, & Stanovich, 1999; Stanovich & West, 1997, 1998a, 2007). To measure general, verbal, and numerical cognitive abilities, the 3-item Cognitive Reflection Test (CRT; Frederick, 2005), the 50-item Wonderlic Personnel Test – Form A (WPT; Wonderlic Inc., 1999), the 10-item Numeracy Scale (NUM; Lipkus, Samsa, & Rimer, 2001), and the 60-item Vocabulary Checklist-with-Foils task (VOC; used as a proxy for cognitive ability in Stanovich & West, 1997) were administered.

2.3. PROCEDURE

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, along with confidence ratings, 2) VOC, 3) CRT, 4) NUM and 5) WPT – Form A. Participants also completed 5 items from the Faith in Intuition subscale of the REI (Epstein, Pacini, Denes-Raj, & Heier, 1996), as well as 5 table-literacy questions adapted from Schield (2005). These additional tasks were not analyzed for the current manuscript. 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.

Table 1. Descriptive Statistics

Category	Measure	Min	Max	Mean	SD	Coefficient alpha
Thinking Dispositions	Need for Cognition (NC)	44	105	73.31	12.58	0.90, as reported by creators of the scale
	Preference for Numerical Information (PNI)	39	120	84.67	15.66	0.94, as reported by creator of the scale

	Actively Open-Minded Thinking (AOT)	124	166	146.13	8.13	Ranging from 0.81 to 0.88, as reported by creators of the scale
Cognitive Ability	Cognitive Reflection Test (CRT)	0	3	1.56	1.17	No psychometric information available in the literature
	Wonderlic Personnel Test (WPT)	14	44	28.41	6.07	Ranging from 0.88 to 0.94, as reported in the user's manual
	Numeracy Scale (NUM)	0	11	9.78	1.63	Ranging from 0.70 to 0.75, as reported by creators of the scale
	Vocabulary Checklist-with-Foils (VOC)	6	33	21.12	4.95	0.87, split-half reliability reported by Stanovich & West (1997)
Statistical Reasoning Assessment	Correct Reasoning (CR)	.15	.94	.60	.16	0.70, test-retest reliability as reported by Garfield (2003)
	Misconceptions (MISC)	.03	.66	.27	.12	0.75, test-retest reliability as reported by Garfield (2003)
	Confidence	2.30	6.00	4.84	.68	n/a

3. RESULTS

The effect of gender and training was analyzed in relation to both performance scales (correct reasoning and misconceptions – Section 3.1), as well as confidence (Section 3.2), using a 2 (gender) × 3 (experience) between-subjects ANOVA. All descriptive statistics are available in Table 2. As mentioned earlier, a high score on the CR scale indicates better performance, whereas a low score on the MISC scale is indicative of better performance (i.e., better avoidance of misconceptions).

The subsequent set of analyses was concerned with the relations among thinking dispositions, cognitive ability, and statistical reasoning. Firstly, zero-order correlations were obtained (Section 3.3). Secondly, the appropriateness of Stanovich's (2009) tri-partite model was tested using confirmatory structural equation modelling (Section 3.4). As appropriate, the role of gender as a predictor was examined. The significance level for all analyses was set to $p < 0.05$.

Table 2. Descriptive Statistics – Performance scales and Confidence (1-6 scale)

	# stats courses	Male			Female			Total		
		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

3.1. EFFECT OF GENDER AND TRAINING ON PERFORMANCE

As mentioned in Section 1.3, we expected that males would outperform females on the SRA and that performance would improve with additional training. Whereas the gender gap could potentially decrease with increased experience, past research does not give a clear indication of whether or not an interaction should be expected. The data were analysed with a 2 (gender) × 3 (experience) analysis of variance, both for the correct reasoning score (CR) and for the misconceptions score (MISC).

As predicted, males performed better than females overall (see Table 2), scoring higher on the CR scale by an average of 12 percentage points across experience levels, $F(1, 193) = 26.31$ $MSE = .020$, $p < .001$, $\eta^2_p = .120$, and committing fewer mistakes, thus scoring lower on the MISC scale by an average of 7 percentage points across experience levels, $F(1, 193) = 16.73$, $MSE = .012$, $p < .001$, $\eta^2_p = .080$. This main effect of gender occurred while effectively controlling for experience in the comparison above. Although it could potentially be argued that increments in performance were simply due to higher general cognitive ability in males, the effect of gender remained statistically significant for correct reasoning ($p < .001$, $\eta^2_p = .227$) even after controlling for intelligence using the well-established WPT as the covariate. However, the effect of gender was no longer significant for misconceptions ($p = .22$, $\eta^2_p = .008$) after controlling for intelligence, potentially indicating that general intelligence plays an important role in avoiding misconceptions.

Increased experience was also associated with better performance (see Table 2), as expected. Indeed, additional courses in statistics were 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 experience level in our analysis performed significantly better 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 training in statistics (*Tukey HSD*, $p = .17$). This latter finding 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, $F(2, 193) = 1.44$, $MSE = .012$, $p = .241$, $\eta^2_p = .015$.

3.2. EFFECT OF GENDER AND TRAINING ON CONFIDENCE

If participants are well calibrated, that is 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) × 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 were more confident than females, $F(1, 193) = 16.91$, $MSE = .379$, $p < .001$, $\eta^2_p = .081$, and increased experience led to greater confidence, $F(2, 193) = 11.57$, $MSE = .379$, $p < .001$, $\eta^2_p = .107$ (see Table 2). Nonetheless, closer examination of the effect of experience revealed a different pattern. Whereas experience continued to have incremental effects on performance with additional courses in statistics, 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 beyond the first course (*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 two or more 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 there is a marginal trend indicating that males' confidence was not as strongly correlated with their performance ($r = .24$, $p = .023$) as females' confidence was with their performance ($r = .48$, $p < .001$); $z = 1.91$, $p = .056$. The main effects of gender on confidence, and the marginal gender effects for the calibration correlations should be interpreted with a note of caution, as follow up experiments that included manipulations of stereotype threat using the SRA and a different statistical assessment task (i.e., the CAOS) did not fully replicate these patterns (see Martin, 2013).

3.3. INDIVIDUAL DIFFERENCES – CORRELATIONS

A first examination of the correlation matrix (see Table 3) revealed that the majority of associations are in the predicted direction, with all measures of individual differences correlating positively with correct reasoning, most measures (except AOT) correlating negatively with misconceptions, and most measures (except AOT and VOC) correlating positively with confidence. 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. The correlation with the Vocabulary task was exceptionally low in comparison to the correlations with the CRT ($r = .56$) 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 assumptions. 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.

Table 3. Correlation Matrix

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

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

3.4. 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.

To examine the relations between thinking dispositions, cognitive ability, and statistical reasoning, a structural equation model (presented in Figure 1) 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). In the current study, thinking dispositions (latent variable) were captured through three measured indicators: the Need for Cognition Scale, the Preference for Numerical Information, and the Actively Open-Minded Scale. Similarly, cognitive ability (also a latent variable) was captured through four measured indicators: the Cognitive Reflection Test, the Wonderlic Personnel Test, the Numeracy Scale, and the Vocabulary-Test-with-Foils. The path model illustrates the relations among the main constructs of interest. In this case, the

path model includes only latent variables (i.e., thinking dispositions (TD), cognitive ability (CA), and statistical reasoning (SR)) and depicts the causal model proposed by Stanovich. 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).

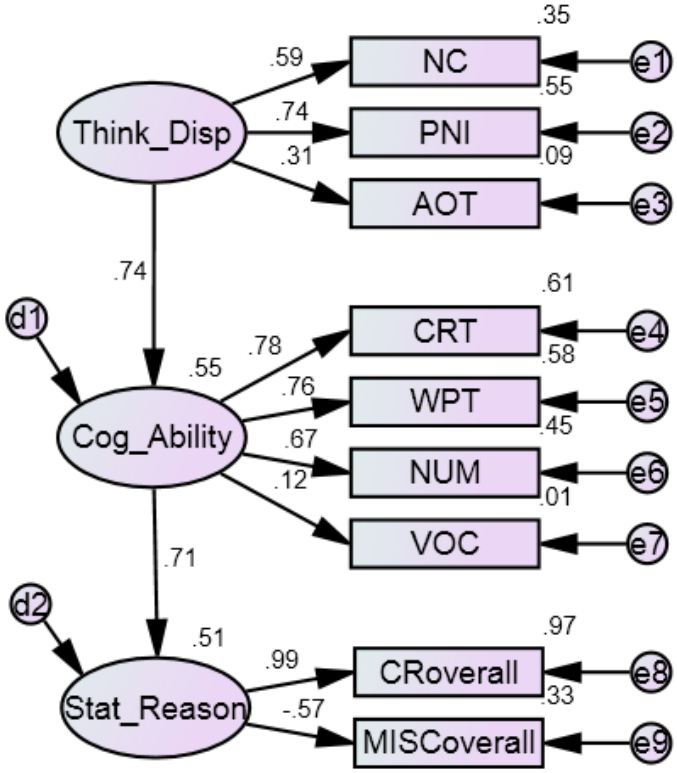

Figure 1. First version of the Structural Equation Model based on Stanovich (2009); standardized estimates are presented

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 is acceptable, the equivalence of the reasoning process across gender will be ascertained. To do so, path coefficients will be 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. The structural equation model 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 & 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 1), complemented by the aforementioned measured indicators, is the basis for the confirmatory test of the proposed model of reasoning. Given that we had no a priori theoretical reason to believe that measurement errors would be correlated in our model, we did not include any covariances between error terms in this model or the models that follow. The model fit indices presented below provide a test of whether such an assumption is reasonable. We did test an additional model that included a latent factor to capture common method variance, with each of the Thinking Dispositions and Cognitive Ability items as indicators. CR and MISC were not included due to issues of model identifiability. Results of this common method variance model did not show any significant factor loadings on the method factor, and this model did not fit the data significantly better than a model excluding the method factor, $\chi^2 = 6.31$, $df = 6$, $p = .39$. Thus, the method factor is not included in the models presented in text.

The first model (see Figure 1) included all indicators for each latent variable. Despite a significant Chi-square test ($\chi^2 = 46.32$, $df = 25$, $p = .006$), 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 2), 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 depicted in Figure 2. 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.

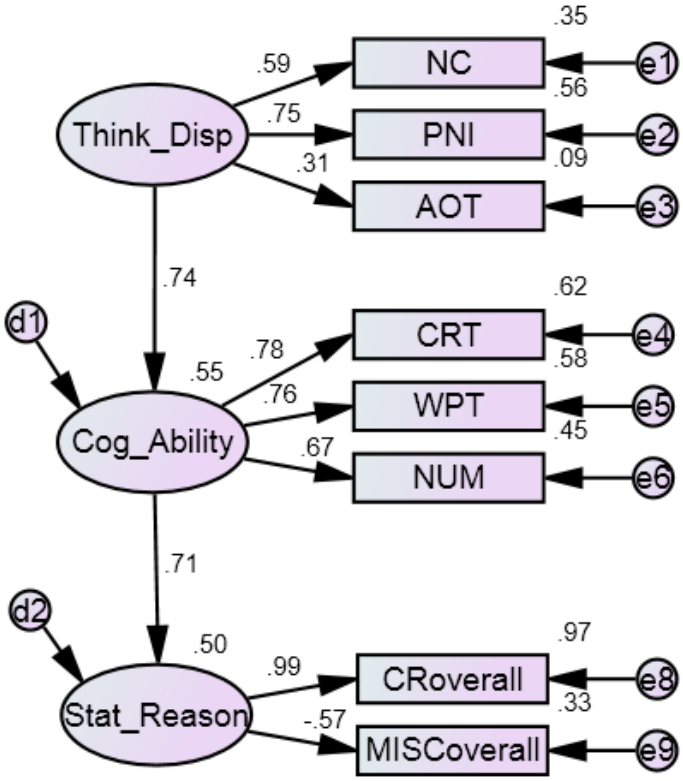

Figure 2. Second version of the Structural Equation Model based on Stanovich (2009). Standardized estimates are presented.

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 are 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 = 42.29, *df* = 38, *p* = .291; *CFI* = .989; *RMSEA* = .024; *pclose* = .887; *ECVI* = .560), 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 model was modified to include this observed categorical predictor variable, with males coded as 0, and females coded as 1 (see Figure 3). The analysis revealed that gender influenced 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), a marginally negative impact on cognitive abilities (-.15, *p* = .065), and a significant negative impact 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 thinking dispositions), and -.16 (39%) remains that is not explained by these two variables.

4. DISCUSSION

In this study, by controlling for experience and individual differences, we provide strong evidence for the existence of a persistent gender gap in statistical reasoning. 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 at least two courses in statistics reached the level of performance of men with no experience in statistics. At the same level of experience, men significantly outperformed them, 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 at least two courses in statistics. This is consistent with prior research by Fong et al. (1986). In their study, participants with one to three 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.

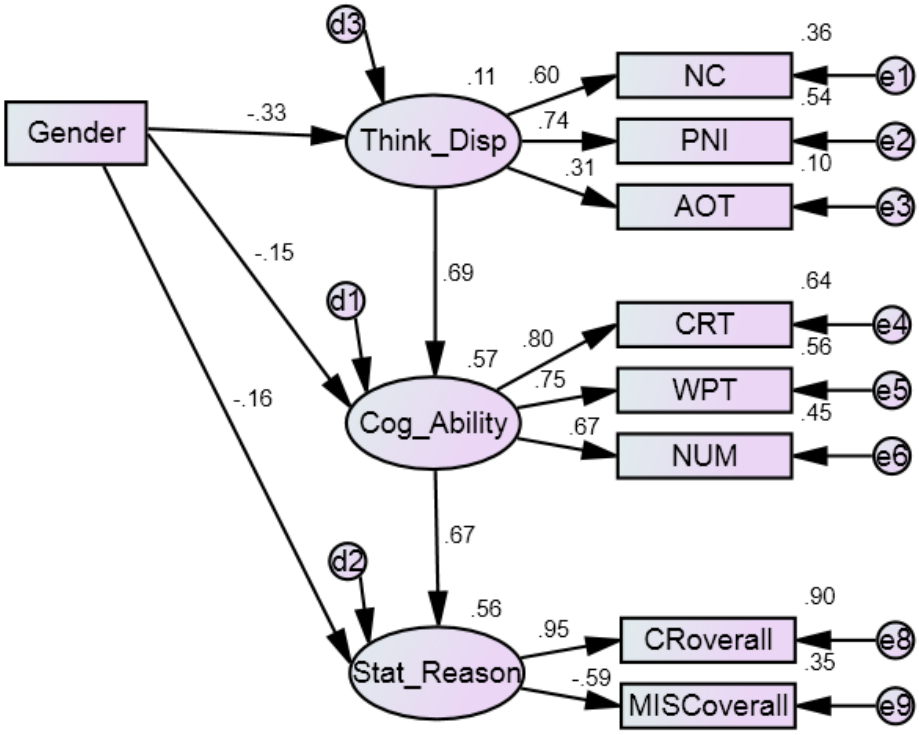

Figure 3. Examining the influence of gender on statistical reasoning. Standardized estimates are presented.

The potential role of confidence (as a potential cause or corollary) of performance is in need of further research. Indeed, the nature of possible interventions would be greatly influenced by an increased understanding of the causal direction of these relationships. In the area of mathematics, the phenomenon of stereotype threat (Quinn & Spencer, 2001; Schmader, Johns, & Barquissau, 2004; Spencer, Steele, & Quinn, 1999), 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 relations 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 influence on thinking dispositions and on cognitive abilities (albeit marginally in the latter case). Indeed, when examining the total effect composed by each of the three 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. Taken together, these results indicate that multiple approaches can be used to attempt to raise the performance of females in statistics, especially via interventions that could raise thinking dispositions 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.

Just as it has been noted in mathematics, multiple factors should be considered when studying gender and performance, ranging from an individual's level of interest in the topic, to cognitive processes, to socialization (Byrnes & Takahira, 1993). In statistics, Gal and Ginsburg (1994) have argued that the achievement of statistical literacy for all must take students' beliefs, attitudes, motivation, and expectations into consideration. It is specifically this need for the integration of cognitive and motivational factors that the current results support, and that researchers and educators in statistics need to take into account. In addition, to continue drawing a clearer picture of gender differences in statistical reasoning, psychological and educational research in statistics could make a point to report systematically the scores of males and females in their research.

It is important to note when considering improving statistical reasoning that 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., 2013), 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.

ACKNOWLEDGEMENTS

Funding for this study was provided by an NSERC Discovery Grant awarded to Jonathan Fugelsang, and an Ontario Graduate Scholarship awarded to Nadia Martin.

REFERENCES

American Statistical Association (2005). *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) *college report*. Alexandria, VA: author. [Online: www.amstat.org/education/gaise/]

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

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, D., & Garfield, J. (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.

Bruner, J. S. (1957). Going beyond the information given. *Contemporary Approaches to Cognition*, *1*, 119–160.

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.

Chance, B. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, *10*(3).
[Online: www.amstat.org/publications/jse/v10n3/chance.html]

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

Cobb, G. W. (1992). Teaching statistics. In L. A. Steen (Ed.), *Heeding the call for change: Suggestions for curricular action*, MAA Notes No. 22, 3–43. Washington: Mathematical Association of American

Couper, M. P., & Singer, E. (2009). The role of numeracy in informed consent for surveys. *Journal of Empirical Research on Human Research Ethics*, *4*(4), 17.

delMas, R. (2002). Statistical literacy, reasoning, and learning: A commentary. *Journal of Statistics Education*, *10*(3).
[Online: http://ww2.amstat.org/publications/jse/v10n3/delmas_discussion.html]

delMas, R., Garfield, J., & Ooms, A. (2005, July). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. In K. Makar (Ed.), *Reasoning about distribution: A collection of research studies. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy* (SRTL-4), University of Auckland, New Zealand, July 2-7, 2005. Brisbane, University of Queensland.
[Online: https://www.causeweb.org/cause/archive/artist/articles/SRTL4_ARTIST.pdf]

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.
[Online: http://iase-web.org/documents/SERJ/SERJ6(2)_delMas.pdf]

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. Freeman and Company.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology, 71*(2), 390–405.

Evans, J. S. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, *128*(6), 978–996.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human perception and performance*, *3*(4), 552.

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

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, *18*(3), 253–292.

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

Gal, I., & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: Towards an assessment framework. *Journal of Statistics Education*, *2*(2), 1–15.
[Online: https://ww2.amstat.org/publications/jse/v2n2/gal.html]

Garfield, J. (1991). Evaluating students' understanding of statistics: Development of the statistical reasoning assessment. In R. G. Underhill (Ed.), *Proceedings of the Thirteenth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 1–7) Blacksburg, VA: VPI and SU.

Garfield, J. (1998) The statistical reasoning assessment: Development and validation of a research tool. In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 781–786). Voorburg, The Netherlands: International Statistical Institute.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, *10*(3).
[Online: www.amstat.org/publications/jse/v10n3/garfield.html]

Garfield, J. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1)*,* 22–38.
[Online: https://iase-web.org/documents/SERJ/SERJ2(1).pdf]

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.
[Online: http://www.jstor.org/stable/1403562]

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.

Jöreskog, K. G., & Sörbom, 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.

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education*, *3*(1), 1–9. [https://ww2.amstat.org/publications/jse/v3n1/konold.html]

Lesser, L. M., Wall, A., Carver, R., Pearl, D. K., Martin, N., Kuiper, S., ..., & Weber III, J. J. (2013). Using fun in the statistics classroom: An exploratory study of college instructors' hesitations and motivations. *Journal of Statistics Education*, *21*(1). [Online: ww2.amstat.org/publications/jse/v21n1/lesser.pdf]

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159–183.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*(1)*,* 37–44.

Lohr, S. (2009, August 5). For today's graduate, just one word: Statistics. *The New York Times*. Retrieved from http://www.nytimes.com

Martin, N. (2013). *Exploring the mechanisms underlying gender differences in statistical reasoning: A multipronged approach* (Unpublished doctoral dissertation). University of Waterloo.

McHugh, R. K., & Behar, E. (2009). Readability of self-report measures of depression and anxiety. *Journal of Consulting and Clinical Psychology*, *77*(6), 1100.

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science*, *238*(4827), 625–631.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*(4), 339.

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.

Parsons, S., & Bynner, J. (2005). *Does numeracy matter more?* London: National Research and Development Centre for Adult Literacy and Numeracy.

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

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

Reyna, V. F., Nelson, W., Han, P., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*(6), 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

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3), 6–13.

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.

Schield, M. (2005). *Five Percentage Table Survey*. W. M. Keck Statistical Literacy Project. Retrieved from 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

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28.

Stanovich, K. E. (1999). Who is rational? Studies of individual differences in reasoning. Mahweh, NJ: Erlbaum.

Stanovich, K.E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J.S.B.T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond*. (pp. 55–88). New York,: 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. (1998a). 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

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

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

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

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.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248.

Wilks, S. S. (1951). Undergraduate statistical education. *Journal of the American Statistical Association*, *46*(253), 1–18.

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

NADIA MARTIN
200 University Avenue West
Waterloo, ON
Canada N2L 3G1