A Sex-Based Approach to Empirical Studies of Programmer Behaviour

Maryanne Fisher
Department of Psychology
Saint Mary's University
Halifax, Nova Scotia, Canada
mlfisher@smu.ca

Anthony Cox
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia, Canada
amcox@cs.dal.ca

Abstract

Sex differences in behaviour have been studied from many perspectives including the psychological and the educational. The literature from these research communities suggests that many of our current practices for obtaining empirical data on reverse engineering may be sex biased and thus negatively impact the measurement of performance. In this short paper, we detail some of the techniques we have used to avoid obtaining sex biased data and to allow more objective scoring of collected data. As side effects, these techniques often improve participation rates and reduce the impact of frequently irrelevant and experimentally confounding interpersonal differences in behaviour.

1 Introduction

It is well known that a person's sex influences their performance on tests and other evaluative procedures. While other variables, such as ethnicity, personality and experience will also influence performance, we focus on sex because there exists a large body of contemporary research on the interaction of sex and performance. We have found that techniques designed to avoid sex-based biases will lead to data that is more representative of individual participants. For example, women's test performance is known to be more impacted by time constraints than is men's performance [7]. Thus, a timed test of programming skill may cause some participants, and more probably women, to perform adversely. However, it is the confound of the time constraint, and not a lack of programming skill, that is the cause of the adverse performance.

It must be stressed that sex differences are not absolute. Instead, they identify areas where there exists a tendency for members of one sex to perform differently to members of the other sex. However, individual variation exists and a specific participant may or may not demonstrate this tendency. Thus, behaviours for which a sex difference exists are often those for which significant individual variation occurs and for which controls must be developed. Knowledge of sex differences will therefore help researchers to avoid situations that create a sex-based advantage and that could prevent accurate measurement of performance.

In the next section, we briefly review some of the key sex differences that can affect the data collected in empirical settings. After this, we examine some of the techniques that we have used to prevent these differences from biasing our data. We then outline some of the approaches that have improved the collection, scoring, and evaluation of our empirical data.

2 Sex Differences

Psychological and educational research has identified several factors that are known to create a difference in performance as a direct consequence of the participant's sex. While many of these differences are significant, it is not possible in this short paper to examine them all. Instead, we focus on attitudes toward risk, a factor that is likely to be the most important sex difference with regard to empirical studies, particularly in areas where one sex is under-represented.

It has been well established that men are more risk prone, or take more risks, than women (see Daly and Wilson [4] for a comprehensive review). Instead, women tend towards a strategy of risk aversion and accept risk only as a last resort, or when the benefits are maximised and costs minimised. In general, women tend to choose high probability, low payoff strategies, whereas men are more likely to choose, under the same situations, low probably but high payoff strategies [4]. This difference in strategy affects experimental performance when participants perceive that there is an element of risk, even of a minimal nature, in their choice of action and their responses to a particular situation or question.

While women are more likely to be more risk averse than men, risk tolerance is an individual preference that will affect any empirical data. Thus, as we advocate through-out this paper, the use of techniques that address a sex dif-

ference will also improve empirical data by avoiding confounding influences introduced by unexpected independent variables. We next examine some of the consequences that can be caused by one's tolerance of perceived risk.

2.1 Problem Solving Strategy

When it comes to problem solving, four distinct approaches have been identified: apply a solution that is familiar, solve it via logical-mathematical reasoning, use trial and error to work backward from a possible solution (i.e., a guess and check approach), and lastly, use a one-shot guess [2]. While all these strategies, and indeed the process of choosing a strategy, have an element of risk, the latter two strategies can be viewed as having higher levels of risk than the first two. Hence, the performance of risk averse individuals is more likely to be impacted when they cannot use the first two strategies. Thus, an interaction occurs between one's risk taking behaviour and tasks that cannot be performed using familiar strategies or logical reasoning.

Standardised testing, such as the Scholastic Assessment Test Mathematics Section (SAT-M), shows that one's sex impacts one's ability to solve conventional problems (i.e., the problem is routine, textbook style or involves the application of known algorithms) and unconventional problems (i.e., uses theoretical mathematics, the application of novel insight, or an unusual use of familiar algorithms). Women are more likely to correctly solve conventional problems using known algorithms, while men are more likely to solve unconventional problems using logical estimation and insight [6]. Byrnes and Takahira [2] found supporting evidence and suggest that for problem-solving, women tend to employ a strategy of applying what is familiar or use trial-and-error.

Furthermore, men tend to take more risks when the need to succeed, or the fear of failure, is intensified [4]. This effect is even greater when only one, high-risk behaviour seems plausible, as other options appear to be dead ends [4]. If an experimenter has developed a study with only one viable solution strategy, men will more rapidly identify a plausible solution and hope that it is "good enough." In contrast, women may feel pressured in these situations and instead take longer to identify the correct solution as a result of evaluating all possible solutions until they are certain of the best response. Women tend to avoid risk and if forced into a risky situation may develop feelings of anxiety, which for computing tasks is deleterious to performance [3].

While these issues particularly affect women, they will affect anyone who prefers to avoid risk. Consequently, experimenters should avoid forcing participants to use an imposed problem solving style. For example, it would be a mistake to require that participants to use a bottom-up program comprehension strategy, unless the experiment's hypotheses are based on the use of this strategy. When a specific strategy is required, time should not be used to measure performance since it is affected by one's tolerance of risk.

2.2 Time Constraints

To minimise risk, women often attempt to conceptualise a potential solution and to spend longer working on this solution before moving on to another problem. Thus, one's tolerance of risk impacts the time one spends on a problem or question and the number of problems that one can attempt. Hence, should a risk averse person take the same test as a more risk tolerant one, the scores may differ because of the care they put into a question and not because of their ability to solve the question.

As an example, the numerous applications of the Vandenburg and Kuse Mental Rotations Test [14] should be considered. In this test, for which a male advantage has been routinely recorded (e.g., Voyer and Saunders [16]), the sex difference has been found to disappear when the time limit for the tests is removed [8]. Thus, we offer evidence that time, and one's perception of its limitations, significantly affects one's performance. While risk tolerance is the likely cause, it is clear that individual differences in tolerance for time limits create differences in performance that can affect empirical results.

Furthermore, as Voyer and Saunders [16] have documented, men are more likely than women to guess a response and men tend to assume that their guesses will be correct. Studies that impose strict time restrictions are thus biased by the participant's likelihood of guessing, and therefore do not accurately ascertain one's true knowledge about an issue. For additional evidence, the findings of Gallagher *et al*. [7] suggest that sex differences with respect to time constraints also exist for standarised testing and is not exclusive to mental rotation ability. They report that on the SAT-M, a timed test, women leave more items blank (unattempted) than do men.

Researchers must therefore control for the effects that time constraints can have on some participants. When a task can not be completed within the time allotted, it may be necessary to reduce the task or increase the allowed time. The use of reaction time when measured unobtrusively will better evaluate performance than will the measurement of completed items within a fixed time interval.

2.3 Performance Anxiety

Females tend to receive the highest evaluations in high-school mathematics classes [11], yet perform more poorly than males on general standardised mathematics tests. One explanation is that the latter environment induces stress and anxiety for some women, and consequently, leads them to

perform poorly [5]. This contention is supported by the idea that standardised tests often lead to the development of stereotype threat; essentially, because women are "expected" to perform poorly on tests involving mathematics (and similarly, those that require programming skills), they will identify with the threat, thus verifying it and performing sub-optimally [5, 13].

Anxiety can be an important influence on performance. Research suggests that the less anxiety an experimental participant feels, the higher their motivation to succeed, and consequently, the better the data because it is more representative of ability [5]. As well, women experience a decrease in self-confidence when they perceive that their work will be compared to others for evaluation [12]. When there is no pressure (i.e., in a non-competitive environment), there is equivalent enthusiasm reported between women and men for computing activities [1]. This finding is supported by data that indicates, for computing tasks, women experience decreased performance in competitive situations [10].

In summary, it can be seen that competition, or the belief that one will be evaluated with respect to others, decreases the performance of some individuals. Experimenters should use techniques that avoid the perception of competition, evaluate participants on a within subjects basis, and provide tasks that are not subject to stereotypic threat.

2.4 Impact of Scoring Procedure

As found by Voyer [15], sex differences in performance are further amplified by the use of negative scoring. That is, if a penalty is imposed for an incorrect response to a test item, women will be less willing to attempt the item when they are uncertain of their solution. When negative scoring is used to dissuade guessing, women are more likely to be influenced and could obtain lower scores. In the study of mental rotation ability, the use of negative scoring has been found to be a significant cause of a sex difference in performance [15]. Experimenters must be aware of this effect and should strive to avoid negative scoring whenever possible. For the most part, techniques that are intended to inhibit guessing are highly impacted by one's confidence. Instead, researchers should seek to identify and measure when participants are unsure of their performance and treat this measure as an additional variable.

2.5 Feedback

Risk aversion, most particularly by women, leads to differences in performance on multi-part tasks. That is, when part B of a task is dependant on the result obtained in part A, one's performance on B is affected by one's confidence in the solution obtained for A. As there is a risk that a mistake has been made, tolerance for risk has a direct impact on the confidence, and therefore the performance, that one will achieve on subsequent elements of multi-part tasks.

Feedback often increases confidence and performance expectations. For example, Lenney [12] found that women have lower expectations for their performance when given ambiguous feedback or no feedback. Research suggests that feedback is particularly important for women as they are more influenced by its content than are men [9]. Thus, researchers must insure that sufficient feedback is available to participants when later elements of a study are dependent on earlier elements. If possible, the elements of a study should be unrelated to avoid the need for feedback.

2.6 Summary of Sex Differences

In summary, we believe that the well researched sex differences in risk perception and acceptance are significant for a variety of reasons. First, accepting risks may be perceived as an indicator of confidence, and perhaps, competence [4]. Thus, in experimental settings, women may be perceived as less competent because they adopt less risky approaches. This matter is even more deleterious, given that stereotypical beliefs, such as "people view me as less competent because I am female," often lead directly to depressed performance [13]. Second, rather than gambling on their ability to guess the correct approach, women will take more time to explore one solution, and will have less time for trying a wider assortment of potential solutions and attempting remaining questions. Third, women will not guess at a response as readily as men and will be less likely to answer questions for which they are not sure of their answers. Multi-part questions that do not provide feedback and negative scoring are additional detriments to the performance of women that will add confounding factors to an experiment.

While we have focused on sex differences, it must be remembered that these differences are driven by risk tolerance, and will affect all participants. Good data collection techniques, as described in the next section, avoid these issues and obtain data that is less influenced by potentially irrelevant factors such as risk tolerance.

3 Improving Empirical Studies

In this section, we examine techniques that we have used to increase the quality of data collected in empirical settings. While these techniques are primarily motivated by our desire to avoid known differences when studying sex-based effects in computer use and programming, we suggest that they will lead to better data for all participants. While trends exist for both men and women, we must stress that a trend is only an increased likelihood of that sex performing a behaviour. Both sexes will exhibit individual differences for a particular behaviour, thus indicating that techniques to

avoid sex differences also help minimise individual differences and avoid unanticipated confounding factors.

3.1 Active, Not Passive, Testing

A wide variety of techniques can be used to collect data. Question-based surveys should be viewed as a "last resort" and should be used when other, more accurate techniques can not be used. Survey questions can lead to a primed response, be misleading because of their wording, or fail to measure an important, confounding variable. They often provide limited context, and thus do not provide participants with sufficient information to truly understand the question or the situation being examined. As well, surveys rely on self-report data, which is highly subject to an individual's perceptions, mood, personality, and level of fatigue. While easy to administer, question-based surveys are often considered boring and may not fully engage participants.

We suggest that an alternative to using question-based surveys is to create more active tasks that address the same theoretical issues. Often, survey questions are used to measure perceptions, attitudes, and specific abilities. In many instances, the same information can be more accurately obtained using alternative techniques that are more engaging and that avoid many of the problems a question-based approach exhibits. From an educational viewpoint, tasks can be viewed as promoting active learning, as opposed to the more passive technique of answering questions.

The techniques used can be simple, easy to create measures. For example, in a recent study, when we asked participants to rank a set of items, we gave participants a set of customised cards to sort. Participants could spread the cards out, view them all simultaneously, and positionally arrange them to match the selected ordering. Participants also indicated that the card sort was fun, novel, and suggested it was easier than trying to number a pre-ordered list of items. We have found that when we use a tactile measure, participants indicate that the task is enjoyable, that they spend longer on it, and that they put more effort into its completion.

In another study, we wished to examine participant's risk tolerance and aversion. While a questionnaire could have been used, we opted to use the Balloon Analog Risk Test (BART). In the BART, participants play a brief computer game in which they earn points for inflating balloons. The balloons explode at some random size, thus examining participants choices with regard to avoiding risk (i.e., the balloons exploding). The BART has been validated in a variety of contexts and is considered highly accurate.

In yet another study, we wished to determine a participant's knowledge of a function's location within a file. Rather than using a Likert-type scale, we gave participants a ruler that represented the file. The participants made a line on the ruler where they thought the method was located. By providing a visual representation of the file, participants had more context than a scale provided, and were not restricted by a scale's granularity.

Many tasks, such as card sorts and the BART, are novel and uncommon and thus avoid the automatic or thoughtless application of learned behaviours, habit-based responses, or stereotypic threat. Tasks (e.g., the BART) often provide more feedback than answering a survey question and therefore decrease effects due to risk tolerance. Tasks are often more tolerant of participant's strategy decisions. A card sort can be accomplished, equally well, using a variety of techniques and provides participants with the opportunity to use the one that is most intuitive or comfortable.

As instructors we do not grade students based on their perception of how well they have learned a topic. Similarly, as experimenters, we must not rely on perceptions and use a variety of techniques to evaluate performance.

3.2 Avoid Abstraction

We believe that the level of abstraction that a task involves increases the perceived risk. That is, concrete tasks generate more immediate and observable feedback while in abstract tasks, one is frequently unsure of correctness until the task is completed. As some participants wish to avoid risk and prefer feedback, the use of highly abstract questions and tasks can decrease the quality, and sometimes prevent the collection, of data from these individuals.

In one study, we asked participants to generate regular expressions for a regular set. In our first attempt, we specified the set using set notation and had many participants not attempt the task. In our second attempt, we embedded set elements within a string and underlined these examples. Compliance significantly increased, thus improving our data collection. Participants suggested that the surrounding text provided context and allowed them to see what was not in the regular set. As well, the approach "felt more real," in the words on one participant. As an earlier task had participants underline the matches to a regular expression within a string, our second attempt asked participants to generate the question, given an answer.

Providing participants with worked examples allows them to test problem solving strategies and generate some initial feedback on their use of these strategies. These characteristics tend to lessen sex differences in performance.

3.3 Accurately Measure Performance

It is crucial that performance be accurately measured. As much as possible, the use of existing, validated techniques is preferred. For example, in recent studies we have used the BART, the Silverman and Eals Object and Location Memory Tests, and the Vandenburg and Kuse Mental

Rotation Test. As these tests have been used by a variety of researchers in a variety of contexts, it was easy for us to find and correct problems in our protocols by comparing our results against those of other researchers. It is notable that, as we previously advocated, all of these tests are task-based.

Many studies on program comprehension and reverse engineering use talk-aloud protocols to determine a participant's focus of attention. We suggest that techniques such as video capture, in conjunction with eye-tracking, lead to better data. In a talk-aloud situation, participants must be constantly reminded to vocalise their thoughts. As they become distracted with the details of a complex task, they frequently forget to vocalise. With eye tracking, it is improbable that participants will forget to move their eyes to the focus of their attention. As well, vocalisations are limited, occur after an action, and are retrospective in nature. Measuring unconscious behaviour, as eye-tracking does, is often more accurate than measuring conscious behaviour, as is done with a talk-aloud protocol.

As well as examining outcomes, measurement of performance should also include the achievement of these outcomes. For example, consider a survey with 20 questions. If participants A and B both have a score of 10, we often assume that they performed equivalently. However, knowing participant A attempted 10 questions and answered all 10 correct while participant B attempted 20 questions and answered only 10 correct, we will likely change our view that they performed equivalently. Using measures from information retrieval, we can more accurately assess performance. Precision measures the number of correct responses with respect to the total number of attempted responses while recall measures the number of correct responses with respect to the total number of possible correct responses. Thus, participant A has a precision of 1.0 and participant B has a precision of .5 while both participants have a recall of .5. As can be seen in this example, by using more detailed measures, a performance difference can be identified and more accurate data interpretation achieved.

4 Conclusion

In this paper, we have presented some of the techniques we have used to avoid sex biased data. However, as sex differences exhibit individual variation, the presented techniques are also effective at limiting the impact of unanticipated individual variation on an experiment. Knowledge of sex-based differences provides researchers with insight into some of the less frequently considered factors that can influence experimental results. The techniques that we have introduced provide researchers with the ability to avoid or decrease the impact that some of these, often undesired, factors can have on empirical studies.

References

[1] O. Astrachan. Non-competitive programming contest questions as a basis for just-in-time teaching. In *ASEE/IEEE Frontiers in Education Conference*, pages T3H/20–T3H/24, Savannah, GA, 2004.

[2] J. Byrnes and S. Takahira. Explaining gender differences on sat-math items. *Developmental Psychology*, 29:805–810, 1993.

[3] J. Cooper and K. Weaver. *Gender and Computers, Understanding the Digital Divide*. Lawrence Erlbaum and Associates, Mahwah, NJ, 2003.

[4] M. Daly and M. Wilson. Risk-taking, intrasexual competition, and homicide. In *Nebraska Symposium on Motivation*, pages 1–36, Omaha, NB, 2001.

[5] J. Duffy, G. Gunther, and L. Walters. Gender and mathematical problem solving. *Sex Roles*, 37:477–494, 1997.

[6] A. Gallagher and R. De Lisi. Gender differences in scholastic aptitude test - mathematics problem solving among high-ability students. *Journal of Educational Psychology*, 86:204–211, 1994.

[7] A. Gallagher, R. De Lisi, P. Holst, A. McGillicuddy-De Lisi, M. Morely, and C. Cahalan. Gender differences in advanced mathematical problem solving. *Journal of Experimental and Child Psychology*, 75:165–190, 2000.

[8] D. Goldstein, D. Haldane, and C. Mitchell. Sex differences in visual-spatial ability: The role of performance factors. *Memory and Cognition*, 18:564–550, 1990.

[9] V. Helgeson. *Psychology of Gender*. Prentice-Hall, Upper Saddle River, NJ, 2^{nd} edition, 2005.

[10] R. Johnson, D. Johnson, and M. Stanne. Effects of cooperative, competitive and individualistic goal structures on computer-aided instruction. *Journal of Educational Psychology*, 77:668–677, 1985.

[11] M. Kimball. A new perspective on women's math achievement. *Psychological Bulletin*, 105:198–214, 1989.

[12] E. Lenney. Women's self-confidence in achievement setting. *Psychological Bulletin*, 84:1–13, 1977.

[13] C. Steele and J. Aronson. Stereotype threat and the intellectual test performance of african americans. *Journal of Personality and Social Psychology*, 69:797–811, 1995.

[14] S. Vandenburg and A. Kuse. Mental rotations: A group test of three-dimensional spatial visualization. *Perception and Motor Skills*, 47:599–604, 1978.

[15] D. Voyer. Scoring procedure, performance factors and magnitude of sex differences in spatial performance. *American Journal of Psychology*, 110:259–276, 1997.

[16] D. Voyer and K. Saunders. Gender differences on the mental rotations test: A factor analysis. *Acta Psychologica*, 117:79–94, 2004.