THESIS

ESTIMATING VARIABILITY ACROSS NUMERIC AND SPATIAL INFORMATION

Submitted by

Kimberly S. Spahr

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2020

Master's Committee:

Advisor: Benjamin A. Clegg
Co-Advisor: Christopher D. Wickens

Mark Prince
Charles Smith

ABSTRACT

ESTIMATING VARIABILITY ACROSS NUMERIC AND SPATIAL INFORMATION

Research has demonstrated the difficulty of estimation and prediction, particularly in complex and uncertain conditions. Specifically, humans lack precision or are biased in making estimates of variability from continuously distributed stimuli, such as hurricane trajectories (spatial information) or sets of random numbers (numeric information). Conversely, people tend to provide calibrated estimates of average behavior for both spatial and numeric stimuli. Using *either* spatial *or* numeric stimuli, past studies noted that people tend to underestimate variability but provide accurate mean estimates. Nonetheless, no experiments have utilized both spatial *and* numeric stimuli to identify the extent to which people estimate variability, and to a lesser extent, mean behavior, across different types of information. This individual differences perspective holds significant implications for training and support in making calibrated decisions under uncertainty.

The current study addressed this gap by presenting participants with a spatial task and a numeric task, each of which assessed knowledge and calibration to variability and means. Using cross-task correlational analyses, this study explored the extent to which similar mechanisms might underlie performance in both domains of stimuli. During the spatial task, participants learned the location of varying trajectories, and then reported on the likelihood of possible trajectory endpoints (spatial variability) and the average trajectory. During the numeric task, participants viewed lists of random numbers, and estimated the mean and spread of these lists (numeric variability). A correlational analysis revealed that participants who gave more accurate estimates of variability on the spatial task were not necessarily more accurate when estimating numeric variability. Such findings indicate that different cognitive processes likely support the understanding of variability for different types of information. Additional research is necessary to elucidate which cognitive mechanisms are involved; possible systems include working memory and numeracy.

Participants expressed a similar overestimation bias to variability across both tasks. This bias trend does not replicate prior literature for either spatial or numeric information, and future studies will focus on how to induce participants to change their response biases. Finally, mean estimation performance correlated across tasks, meaning that those who were more accurate when estimating spatial means were more likely to accurately estimate numeric means.

ACKNOWLEDGEMENTS

I sincerely thank my advisors, Benjamin Clegg and Chris Wickens, for their guidance and support during this project. Further, I deeply appreciate the time and effort of Charles Smith and Mark Prince for being a part of my thesis committee.

TABLE OF CONTENTS

ABSTRACT ii
ACKNOWLEDGEMENTS iv
CHAPTER 1 - INTRODUCTION 1
    Variability, Means, or Both? 3
    Significance of the Current Study 3
CHAPTER 2 – BACKGROUND LITERATURE 6
    Spatial Stimuli 6
        Estimating spatial variability. 6
        Estimating spatial means. 10
    Numeric Stimuli 11
        Estimating numeric variability. 11
        Estimating numeric means. 14
    MOVE: A Model Of Variability Estimation 15
        Cognitive processes of variability estimation. 16
            Encoding. 16
            Retention interval. 16
            Biases. 16
            Response methodology. 17
        Quantifying responses to variability. 17
    Individual Differences 18
CHAPTER 3 – PURPOSE AND RESEARCH QUESTIONS 21
CHAPTER 4 – GENERAL METHODS 23
    Power Analysis 23
    Participants 24
    Analysis Plan 24
CHAPTER 5 – METHODS: SPUN 25
    Phase 1 25
        Phase 1 (mean and variability prediction). 25
        Phase 1 feedback. 26
        Phase 1 variables. 26
    Phase 2 27
        Phase 2 (variability prediction). 27
        Phase 2 variables. 27
        Financial incentives. 28
    SPUN Trajectories 28
    SPUN Methods Summary 29
CHAPTER 6 - RESULTS: SPUN 31
    Variability Estimation 31
        Distribution of responses: Circle size. 31
        Phase 1 variability sensitivity: Circle size. 32
            Degree of sensitivity. 33
        Phase 1 variability bias. 33

Absolute Phase 1 variability error. 34
Distribution of responses: Estimated variability. 35
Phase 2 variability sensitivity: Estimated percentage within the circle. 36
Phase 2 variability assessment and task order. 37
Degree of sensitivity. 37
Phase 2 variability bias. 38
Phase 2 variability bias and task order. 39
Absolute Phase 2 variability error. 39
Mean Estimation 40
Distribution of responses: Mean prediction error. 40
Spatial mean prediction error. 41
Mean prediction and task order. 42
CHAPTER 7 - DISCUSSION: SPUN 44
Task Order 44
Sensitivity to Spatial Variability 44
Speed and angle sensitivity. 45
Variability Bias 46
Estimating Spatial Means 49
CHAPTER 8 - METHODS: DIGIVAR 51
Mean Estimation 51
Variability Estimation 52
Key Variables 52
Number Lists 53
CHAPTER 9 - RESULTS: DIGIVAR 54
Variability Estimation 54
Distribution of responses: Numeric variability. 54
Numeric variability sensitivity. 54
Numeric variability bias. 55
Variability estimates by block. 57
Numeric variability bias and task order. 57
Absolute numeric variability error. 58
Mean Estimation 59
Distribution of responses: Numeric mean error. 59
Absolute numeric mean error. 60
CHAPTER 10 - DISCUSSION: DIGIVAR 62
Task Order 62
Sensitivity to Numeric Variability 62
Variability Bias 63
Estimating Numeric Means 65
CHAPTER 11 - METHODS: CORRELATIONAL ANALYSES 67
CHAPTER 12 - RESULTS: CORRELATIONAL ANALYSES 68
Variability Sensitivity 68
Variability Bias 68
Estimation Error of the Mean 68
Estimation Error of the Mean and Variability Sensitivity 69
CHAPTER 13 - DISCUSSION: CORRELATIONAL ANALYSES 70

Variability Sensitivity 70
Variability Bias 70
Estimation Error of the Mean 71
Estimation Error of the Mean and Variability Sensitivity 72
CHAPTER 14 – GENERAL DISCUSSION 73
Individual Differences 75
Estimation bias. 75
Estimation error of the mean. 76
Observed Power Analysis 77
Limitations and Future Directions 78
Levels of manipulated variability. 78
Financial incentives. 79
Latent model of variability estimation. 79
Refining DigiVar. 80
CONCLUSION 81
REFERENCES 83
APPENDIX A 89

CHAPTER 1 - INTRODUCTION

Humans constantly make "best guess" judgements or decisions with limited information. Take, for example, the outcome of the Kentucky Derby. This racing event draws viewers worldwide, with audiences wagering a large amount of money on the outcome. For example, viewers wagered approximately $138 million on the lineup in 2016 (Isidore, 2016). Given that the event outcome is far from guaranteed, viewers rely on limited cues to decide which horse will come out on top. Even the best gamblers understand that such situations include significant uncertainty: the weather, track conditions, injuries, prior wins, starting positions, jockey choice, and temperament all play some role in whether a horse will win such a race.

Gambling is only one example of high-stakes decisions based on limited information. Weather forecasters make predictions based on a limited number of cues, and those expert predictions are assessed based on their accuracy (i.e., did the predicted weather come to pass?). Actuarial weather judgements come largely from climate models and mathematics (Tyszka & Zielonka, 2002), and while prediction error has decreased significantly with the advent of new technologies and probabilistic models, uncertainty in prediction still remains an issue (Alley, Emanuel, & Zhang, 2019). For example, the rate of error for predicting tropical storms in the Atlantic for the 2015 hurricane season ranged between about 45 nautical miles (n mi) for a 24-hour forecast and 340 n mi for a 120-hour forecast (National Oceanic and Atmospheric Administration [NOAA], 2017). While this forecast error isn't ostensibly alarming, particularly for shorter time intervals, the overall time-accuracy tradeoff can lead to serious safety implications. If an initial 120-hour forecast led people to think that they are safely outside a hurricane trajectory, they may not have time to evacuate come the 24-hour forecast, which shows them right in the path of the oncoming storm. These domains are just two examples of many areas in which humans must make judgements in the face of environmental uncertainty, and due to the often high-stakes outcomes within such judgements, it is important to understand the cognitive processes at play when making these decisions.

Kahneman (2011), Silver (2012), Tetlock (2005), and others fueled academic interest in the process and accuracy of human prediction under environmental uncertainty (i.e., when exhaustive data are not available). Indeed, the need for, and application of, prediction presents itself on a daily basis, ranging from simple tasks with mild consequences (e.g., at what time will traffic on a route dissipate?) to major decisions with global impact (e.g., should New Orleans evacuate due to a severe hurricane?). Generally, predictions are hard, and humans often demonstrate overconfidence in the accuracy of their predictions (Einhorn & Hogarth, 1978; Kahneman, 2011). Specifically, regarding the prediction of continuous dynamic trends (e.g., changes in economic indicators, the trend of a hurricane, or uncertain trajectory of an aircraft), humans tend to accurately identify average trends, but underestimate trend variability (Herdener, Wickens, Clegg & Smith, 2016; Pugh, Wickens, Herdener, Clegg, & Smith, 2018). Prior research supports this generalization; studies focused on estimating average behavior are very consistent, concluding that typically, such estimates are precise, regardless of the stimuli. Conclusions about variability estimation, however, are they, themselves, variable. Less is known about how people estimate variability, and the extent to which estimation performance is the same across different stimuli (e.g., spatial versus numeric trends).

Research on how humans estimate characteristics of continuously distributed stimuli (Hofstatter, 1939; Beach & Swenson, 1966; Beach and Scopp, 1968; Peterson and Beach, 1967; Lathrop; 1967; Herdener et al., 2016; Herdener et al., 2017; Pugh et al., 2018; Herdener et al., 2018; Herdener et al., 2019) demonstrate that the type of information presented may impact responses. Primarily, though, this literature does address how people estimate trends across different types of stimuli; participants in these previous studies have either worked with spatial or numeric stimuli, but not both. No one has yet identified how the same participants estimate variability and means across *more than one* type of information – for example, if one is accurate when estimating the variability of spatial trajectories, are they also accurate when estimating the variability in a number list?

Variability, Means, or Both?

This project is part of a series on human responses to environmental uncertainty. The paradigms we developed to study responses to uncertainty include measures of mean and variability. While responses to both distribution characteristics are consequently examined here, our dominant focus is on variability, given that it most closely corresponds to uncertainty (i.e., presenting variable information naturally conveys a level of uncertainty). Furthermore, existing studies suggest that it is still unclear how people estimate variability and if performance changes based on the stimuli used. Conversely, trends in mean estimation are much more consistent, and need little new empirical attention. Thus, hereafter, the emphasis will be placed on how humans estimate variability.

Significance of the Current Study

Understanding the extent to which people respond similarly to spatial and numeric variability has clear implications for both theory and application. Regarding theoretical implications, the current study represents a broadening of perspective compared to other literature in this area using an individual differences approach. As mentioned above, most studies only focused on one type of stimuli, so conclusions about the underlying processes are restrictive. This is a major limitation for creating overarching cognitive theories about how humans understand uncertainty.

Cognitive abilities are commonly described as either unidimensional/global or multidimensional/specific, and our understanding of these cognitive processes oscillates between the two perspectives. As one example, the concept of working memory experienced a similar shift, moving from a single storage mechanism in Shiffrin and Atkinson's (1969) model of memory to the active multifaceted system in Baddeley's (2001) model, responsible for information integration. Opposite trends occur as well, with researchers demonstrating that concepts previously considered completely different actually have shared variance in performance, like working memory capacity and attention resources (Engle, 2018). New ways of assessing these major cognitive constructs led to significant reorganization of their research structure (Matthews, Davies, Westerman, & Stammers, 2013). This type of empirical evolution alters the nature of subsequent predictions and overarching theories about cognition.

The pre-existing literature on variability in particular is largely disparate, having participants estimate characteristics of only one type of stimuli using different response methods. Studies that utilize only one type of stimuli may note accurate trends, but they are consequently domain specific. For example, Herdener and colleagues (2016) found clear evidence that people underestimate spatial variability (i.e., responses to variability are much smaller than the true level of variability). While this robust trend speaks to how people understand varied spatial information, one cannot draw any conclusions about how people process other types of variability (e.g., numeric). Having participants estimate characteristics of both spatial and numeric stimuli will address whether humans have *global* abilities to understand variable behavior.

Changes in theory like the examples above have a domino effect when applying research outside of the laboratory. For example, human factors psychology focuses on how to best utilize research and improve human performance (Wickens, Hollands, Banbury, & Parasuraman, 2013), commonly through training and visualizations. If the understanding of average and variable behavior is a unitary cognitive process, then having people train to make better judgements with spatial stimuli should lead to benefits when making that same judgement with numeric stimuli. Conversely, if cognitive comprehension changes with different stimuli, then training must be done using both types of stimuli.

Thus, the current study addresses this gap by focusing on the individual differences in how people estimate variability, specifically identifying the extent to which absolute error/biases expressed when estimating from spatial stimuli will be similar to those seen with numeric stimuli. The following sections provide a foundation of relevant background literature regarding the estimation of 1) Spatial variability, 2) Spatial means, 3) Numeric variability, and 4) Numeric means.

CHAPTER 2 – BACKGROUND LITERATURE

Spatial Stimuli

Making predictions using spatial stimuli (i.e., non-numeric information) is necessary for a number of daily activities. For example, when choosing where to merge into traffic, one must estimate the proper space between cars in an adjacent lane. Viewing weather trajectories on a map is another example of spatial information from which people make predictions. This information cannot easily be represented by numbers, but one can still describe its behavior in terms of variability and means. Empirical trends suggest that people struggle when estimating spatial variability, demonstrating significant underestimation compared to the true values (Herdener et al., 2016) and decision-making heuristics may influence this underestimation. Conversely, people seem to accurately predict spatial means (Herdener et al., 2016).

Estimating spatial variability. The variability in a spatial distribution conveys important information; for example, in a hurricane "cone of uncertainty," the visualization only encompasses two-thirds of past forecast modeling errors (NOAA, 2019), meaning that there is a one-third chance that the storm will fall outside the cone. Clearly, proper understanding of spatial variability is an important area of study to change public behavior.

Early studies that probed the cognitive understanding of spatial variability focused on how the average of a distribution influences variability estimates. Using bundles of sticks with variable lengths, Hofstatter (1939) had participants make judgements about the variability of lengths in each bundle. These judgements increased as the true variability increased, as expected, demonstrating sensitivity to changing variability. Interestingly, though, when the mean of the distribution increased, variability judgements were suppressed. This suggests that participants underestimated variability when the distribution had a larger mean (Brunswick, 1956). Importantly, this early study showed that people were not blind to changes in true variability, but also that the magnitude of the distribution mean influenced performance when estimating that variability.

Peterson and Beach (1967) interpreted this finding, suggesting that the distribution mean exerted a suppressive influence on variability judgements, a phenomenon that they compared to the Weber Fraction. The Weber Fraction, used primarily in psychophysics, suggests that the perceived change in a stimulus follows a ratio with the initial intensity of that stimulus, where incremental intensity judgements are lower when the initial stimulus starts higher (Norwich, 1987). In a bright room, it is harder to perceive an incremental change in the intensity of lighting compared to being in a dark room. The initial stimulus intensity serves as an anchor, and when that anchor starts high, people struggle to notice subsequent small adjustments. Applied to Hofstatter (1939), the distribution's *mean anchor* disproportionately influenced participants to underestimate variability. This phenomenon is the same as Tversky and Kahneman's (1974) anchoring and adjustment heuristic, wherein the conclusions drawn from information (i.e., variability judgement) favor the initial, and often strongest, data (i.e., mean).

Lathrop (1967) replicated the Hofstatter (1939) study using cards that contained distributions of lines with different length means, variability, and sequencing. Participants responded by ranking the distributions in order by variability. Echoing the results above, distributions with a larger mean were thought to have a lower overall variability. Moreover, the sequence of lines within the distributions also influenced perceived variability. Specifically, variability judgements were most accurate as the sequence of lines increased and decreased, as in a normal distribution. Peterson and Beach (1967) pointed out that the sequencing effect works as follows: to make judgements about variability, participants must assess the deviation of each instance from the group mean, and when these discrepancies are large, people tend to underestimate the overall variability. In a normal distribution, most values are positioned close to the mean, overall deviations from the mean are small, and thus, variability judgements are generally accurate. When distributions are non-normal and contain many extreme values (e.g., a saddle-shaped distribution), participants would be more likely to underestimate variability, as demonstrated by Lathrop (1967). Conclusions from Hofstatter (1939) and Peterson and Beach (1967) could work in synergy: if participants cannot perceive smaller deviations when the distribution has a larger mean, then they may focus more on the extreme values. As seen in Lathrop (1967), this often leads to the underestimation of variability.

Using a paradigm that presents numerous, variable hurricane track-like spatial trends, Herdener and colleagues (2016) also assessed the understanding of variability in two different ways, with similar results to those studies discussed above. In the paradigm, participants would see the location of trajectories at two time points (e.g., T0, T1), representing movement through space. They were then asked to predict the most likely location of T3 and size a circle that captured 70% of the possible trajectory endpoints, based on the distribution of all prior trajectories (Herdener et al., 2018, 2019; see *Figure 1*). After viewing and predicting sets of 20 spatial trends, participants were asked to provide numeric point estimates of variability (e.g., 50%, 70%) at different points within the distribution. These variability response methods are perceptual adjustment and numeric assessment, respectfully. Significant underestimation of variability occurred with both types of response probes. Underestimation was more extreme, however, for the numeric estimates compared to the spatial adjustments. Implications of this finding will be discussed below in the theoretical framework section.

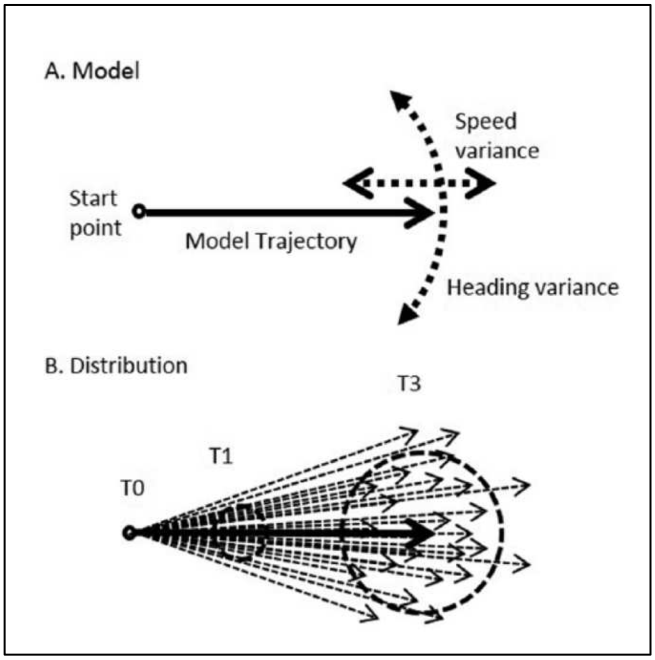

Figure 1. SPUN Diagram. This figure presents a schematic diagram the SPUN paradigm which has participants predict the most likely location of T3 in a trajectory and then size a circle that captures 70% of the possible trajectory endpoints.

Taken together, the above studies suggest that people do not make unbiased judgements of spatial variability; indeed, the trend of underestimation was robustly noted, and was often influenced by the mean and sequence of stimuli. Importantly, this response bias occurred via three different types of variability probes: point estimates (Hofstatter, 1939; Herdener et al., 2016, 2018), "more or less variable" judgements (Lathrop, 1967), and perceptual adjustment (Herdener et al., 2018). This means that these results are not just an artifact of the response methods used; people show a pervasive bias to underestimate variability.

Kareev, Arnon, & Horowitz-Ziegler (2002) examined *why* people underestimate the variability of a distribution. This group posited that when people make inferences about a population's variability, they use a small, biased sample of information. Given that variability in the population is larger than that of a sample, if participants use a sample of information, they naturally underestimate the true variability. Indeed, this is why we apply a correction when calculating sample statistics (unlike population statistics). For example, a sample variance requires division by *N* - 1, as opposed to the population variance, which requires division by the full sample size *N*. Kareev, Arnon, & Horowitz-Ziegler point out that, people, when making estimates about the population, do not intuitively correct for this downward-biased phenomenon, resulting in the empirical trends seen above. Indeed, in a series of experiments with non-numeric stimuli, they again found underestimation of the population variability based on how participants sampled information.

This explanation suggests that responses to variability occur via the representative heuristic, or cognitive shortcut, often used when situations are uncertain (Tversky & Kahneman, 1973). Given a sample of information, participants respond to variability as if that sample perfectly represented the population of interest. Doing so can evoke an underestimation bias when making such judgements (Hansson, Juslin, & Winman, 2007).

Estimating spatial means. Prior research suggests that people can accurately predict the mean from a series of spatial trajectories. Using the same paradigm described above, Herdener and colleagues (2016) asked participants to learn the average behavior of continuously presented spatial trends. Again, after viewing the location of trajectories at two time points (e.g., T0, T1), participants were then asked to predict the most likely location of T3. This response requires an understanding and extrapolation of average trend behavior (see *Figure 1*). In Herdener et al., (2016), these estimated locations were compared to the true trajectory model means. Importantly, for each successive mean estimate, participants were given feedback on the true trajectory mean, providing an opportunity to learn average behavior.

These researchers found that participants could accurately predict a mean trajectory after four or five trials within a block. These mean estimates were more accurate for linear trajectories compared to curvilinear trajectories (Pugh et al., 2018). Overall, people rapidly oriented to spatial mean information and adopted an optimal strategy to predict such an average. These findings have been consistently replicated (Herdener et al., 2018, 2019).

To summarize, a variety of research has examined how people estimate numeric means and variability, leading to the following key trends:

A. Distribution characteristics (i.e., mean, sequence) influence estimates of variability;
B. Participants consistently underestimate spatial variability. This occurs with different response methods;
C. Biases in variability judgements may come from decision making heuristics; and
D. Participants tend to give accurate estimates of the mean of spatial trajectories.

Are these trends unique to spatial stimuli, or do they generalize to other types of information? As discussed next, similar performance patterns have been observed with numeric stimuli.

Numeric Stimuli

Humans are surrounded by numbers. For example, people track the rise and fall of the stock market. Others weigh themselves each morning and track the weight they lose over time. Meteorologists quantify the likelihood of temperatures and laypeople use this information to dress appropriately. Due to the abundance of numeric information in the world, a large body of research has focused on how people estimate the characteristics of number sets, including variability and means.

Estimating numeric variability. As discussed above, people struggle to accurately estimate numeric variability (Pollard, 1984). Beach and Scopp (1968) demonstrated this observation using four decks of cards, each containing 20 numbers. The decks had the same mean, but different variances. Participants viewed pairs of decks, with the numbers presented sequentially, and were asked to state which deck had the higher variance. After making these judgements, half of the participants returned and viewed the same decks, this time to estimate the ratio of the larger variance to the smaller variance. Results suggest that, while participants accurately judged which deck had the higher variance, their ratio responses were systematically smaller than the true ratios. This means that they judged the variability of the larger deck much smaller than it truly was, relative to the variability of the smaller deck. This trend shows that, while people demonstrate sensitivity to higher versus lower variability, they still tend to underestimate the magnitude of that variability. Furthermore, this estimated ratio did not increase linearly with the true ratio, indicating that participants were less than fully sensitive to changes in variability

In a similar study, Henrion & Fischoff (1985) examined how scientists represented error in measurement, or systematic variability, in their estimates of physical constructs (e.g., the speed of light, Avogadro's Number). These researchers point out that estimating such constants is akin to making judgements under uncertainty; consequently, they hypothesized that even highly trained scientists demonstrate an underestimation of variability surrounding their numeric predictions. Such measurement error around a numeric point estimate (i.e., a confidence interval) conceptualizes an estimate of numeric variability. Henrion & Fischoff (1985) retrospectively examined how both point estimates and confidence intervals of physical constructs changed from 1958 to 1973. As one would expect, across time, measurement accuracy of point estimates increased, but interestingly, the confidence intervals around each point estimate were small. Often, when a construct was measured, re-evaluated, and published with a new "best-guess" estimate, the new value was *well* outside of the previous measurement's confidence interval, meaning that the confidence interval for the initial estimate was likely too small.

Henrion & Fischoff (1985) interpreted this finding as overconfidence in scientific estimates and associated confidence intervals. In terms of variability, too-narrow confidence intervals suggest the people underestimated the numeric variability surrounding their measurements. While the authors suggested that the small intervals could have resulted from the, "…difficulty of thinking of reasons why one's best guess might be wrong (pg. 796)," perhaps the phenomenon is better explained in terms of Tversky & Kahneman's (1974) anchoring and adjustment hypothesis: these scientists anchor on their point estimate and do not adjust their interval sufficiently enough to accommodate all possible sources of error. In sum, this key study demonstrated that even highly trained individuals tend to underestimate numeric variability.

In another study concerning numeric intervals, Laestadius (1970) showed their participants sets of numbers and asked for estimates of confidence intervals around a mean. Some of these number lists had low variability and others had high variability. No biases were assessed in this study, but results indicated that intervals for high variability lists were significantly larger than those for the low variability lists, suggesting some sensitivity to increasing variability.

Finally, Hansson, Juslin, & Winman (2008) created an experiment during which participants estimated the mean income of various fictitious countries with feedback on these estimates. After estimating the means, participants were given the true mean income for other countries and were asked to provide numeric confidence intervals around those new provided means. In three different experiments, results demonstrated that the participant-generated intervals were grossly smaller than the true intervals. Such findings correspond to those discussed above: people tend to underestimate numeric variability.

While few studies have examined how people estimate numeric variability, the results presented above clearly demonstrate that people show some sensitivity to increasing variability, and that often they underestimate that variability.

Estimating numeric means. A similar parallel occurs between research on the estimation of numeric and spatial means; people tend to provide accurate estimates of numeric means. Further, across multiple studies, characteristics of numeric distributions can influence these judgements. Spencer (1961) published the earliest study examining how people intuitively estimated numeric averages. In this study, people accurately estimated average values, but error increased when the variability of numbers around the mean also increased. This trend suggests that some information not directly relevant to the mean can nonetheless influence mean judgments. Beach and Swenson (1966) corroborated these findings using lists of 3, 5, and 7 numbers. Their participants estimated numeric means with the low average estimation error of 3.51 digits, but overall error increased as the variability between numbers increased. These findings suggest that, as numbers further away from a list mean are presented to participants (i.e., the list had higher variability), it is more difficult to weight the deviations, and hence, accurately estimate the mean. No consistent directional biases occurred: in two of the three mean estimation conditions, participants slightly underestimated the mean, and in the final condition, slightly overestimated the mean. In sum, these studies suggest that people can often reliably estimate a mean, but error increases as list variability increases.

The Laestadius (1970) study introduced above also examined how list variability influenced participant generated mean estimates. This study found that high variability number lists elicited more error in mean estimates compared to the low variability lists, echoing Spencer (1961) and Beach and Swenson (1966).

Overall, these studies demonstrate that, while people accurately estimate numeric means, absolute error increased as variability increased. This means that there is a link between numeric variability and estimating a mean, but none of these studies clearly discussed the bias expressed by participants. This missing aspect of performance will be examined in the current study.

In conclusion, the studies discussed above include the following general observations:

A. Participants demonstrate sensitivity to higher or lower numeric variability, but they still underestimate that variability. This underestimation also occurs with highly trained scientists;

B. Variability underestimation may occur due to anchoring on a point estimate, or mean, and insufficiently adjusting their interval when estimating variability;

C. People tend to be accurate when they estimate numeric means, but this accuracy decreases as variability increases.

The summary points for spatial and numeric stimuli are consistent, but this has not been verified through a within-subjects design. Examining judgements across stimuli will identify the extent to which similar cognitive processes govern the understanding of variability (and to a lesser extent, means). The current study addresses this literature gap, using the Model of Variability Estimation (MOVE) as a theoretical framework.

MOVE: A Model Of Variability Estimation

Wickens, Clegg, Witt, Smith, Herdener, and Spahr (2020) presents a theoretical framework of variability estimation, tying together many of the findings presented above. This model describes the cognitive processes involved in variability estimation, suggests estimation performance metrics, and provides a structure for interpreting findings from studies focused on variability.

Cognitive processes of variability estimation. MOVE argues that four main cognitive processes facilitate the creation of a mental model of variability within a set of continuous stimuli. This mental model, in turn, influences responses to variability.

Encoding. When people create a mental model of variability, they must have been exposed to multiple instances of continuously distributed stimuli. This exposure, referred to as encoding, is shaped by experimental variables such as presentation speed, familiarity of stimuli, attention paid to the stimuli, and stimuli order.

Retention interval. Retention interval, or the time between encoding and an individual's response, also impacts performance. Without active rehearsal, information decays from short term memory quickly, so with a longer retention interval, responses to variability will be less accurate.

Biases. These phenomena describe systematic errors that occur during encoding and/or the retention interval. Examples of biases at play here include anchoring and adjustment (Tversky & Kahneman, 1974), and primacy or recency effects. Primacy refers to remembering the first instances of a set better compared to others in the set. Recency, conversely, refers to remembering the final instances of a set better than other instances. Prior research has hypothesized about how such biases influence responses to variability. For example, Herdener et al. (2018) found that the mean of a spatial distribution functioned as an anchor and influenced judgements of variability. In another experiment, participants drew samples from a distribution of colored beads, and subsequently predicted the variability of the population. In this study, participants overwhelmingly based their variability predictions on the first ten items they saw, suggesting a primacy effect (Kareev, Arnon, Horowitz-Zeliger, 2002).

Response methodology. Research suggests that the probes used for people to respond to variability will influence performance as well. As described above, our lab has used both *adjustment* of a confidence interval (i.e., a circle to encompass the variability of encountered instances) and *estimation* of variability (i.e., "what percentage of the stimuli you just saw fell within this circle of fixed size?"). We have generally found that adjustment produces less bias (either over- or underestimation) than estimation (Herdener et al., 2019). In this same study, participants also demonstrated more sensitivity to increasing variability using adjustment compared to estimation.

Quantifying responses to variability. To a certain extent, how people understand variability is based upon how we quantify participants' responses. MOVE describes the following three ways to do so: 1) sensitivity, 2) bias, and 3) precision (absolute estimation error). Each computation captures unique information about responses to variability (see *Figure 2*).

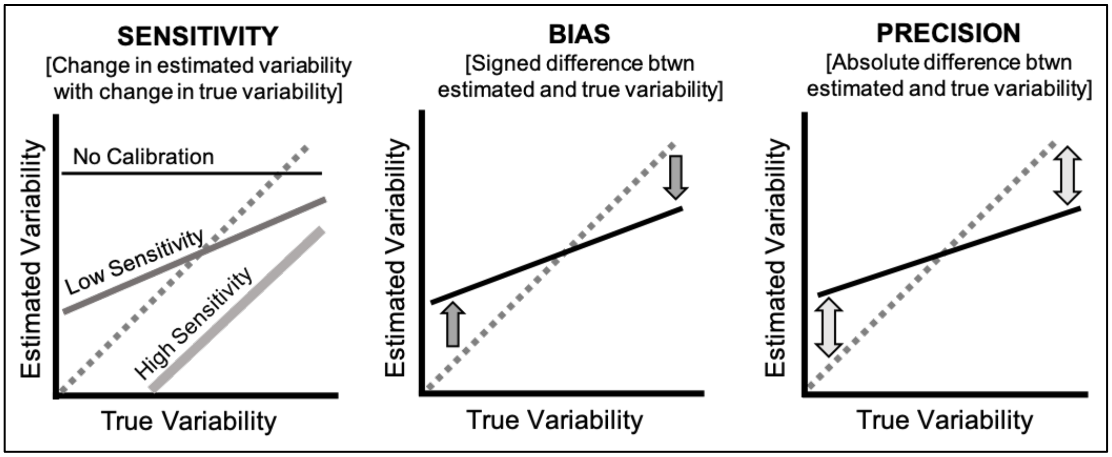

Figure 2. Quantifying responses to variability in the MOVE Model. This diagram presents three methods to assess responses to variability, including sensitivity, bias, and precision.

In these diagrams, estimated variability is plotted as a function of true probability, and this linear function can be compared to the perfect calibration dashed diagonal line (i.e., representing perfect responses – as true variability increases by a set amount, estimated variability increases by that same amount).

Represented by the slope of the response line, sensitivity is essentially a psychophysical function that refers to how well people, shown different sets of stimuli whose variability differs, can notice these differences (see the leftmost panel in *Figure 2*). This slope matters, because if participants are blind to variability (top line), it makes any bias expressed less meaningful as the entire measure of bias will depend only on the level of true variability presented. The positive slope of the low sensitivity line demonstrates some sensitivity to changes in true variability, but less than optimal. We classify the high sensitivity line as well-calibrated because it has a nearly identical slope compared to the perfect calibration dashed line.

Bias, or the signed difference between true and estimated variability, represents the extent to which participants over- or underestimate variability (see the middle panel in *Figure 2*). Variability responses above the dashed calibration line correspond to overestimation, and values under the dashed line correspond to underestimation. Finally, precision refers to the absolute error for estimated variability (see the rightmost panel in *Figure 2*). While MOVE uses the term "precision", from here onward, this study will use the term "absolute error" for ease of interpretation. The performance metrics of sensitivity, bias, and absolute error will be utilized throughout this study.

Individual Differences

People vary on ability and performance. Individual differences studies investigate these variations in performance to understand, predict, and create theories as to why this occurs

(Sternberg, 1999, Matthews et al., 2013). Individual differences in cognitive abilities may account for the empirical variability trends noted above.

Some interpret the general tendency to underestimate variability as overconfidence in the estimation of the mean (Henrion & Fischoff, 1985; Wickens et al., 2020). As estimates of variability represent environmental uncertainty, providing an overtly small estimate regarding the variability in a distribution suggests that people have a larger-than-warranted confidence in their mental model. Using an individual differences factor analysis, Pallier and colleagues (2002) found evidence of a confidence trait which correlated with other general cognitive abilities. For example, those with lower intelligence (captured via cognitive testing) tended to express more overconfidence. Further, both the traits of proactiveness and activity positively correlated with confidence biases, suggesting a relationship between specific personality traits and expression of overconfidence. Experimenters did not assess estimation of variability in this factor analysis, but nonetheless, they did demonstrate individual differences in the expression of confidence biases. To extend this finding, as underestimating variability corresponds to overconfidence, individual differences in overconfidence could also motivate estimates of variability. Individual differences in variability estimations have been concretely addressed by a few studies.

Kareev, Arnon, & Horwitz-Zeliger (2002) tested differences in participants' working memory capacity (WMC) and correlated those results with predictions about the variability in the composition of an unknown population. They found evidence that variability predictions were correlated with the amount of information considered (i.e., sample), and how much information people could consider correlated with WMC. In this way, differences in WMC, a cognitive "ability" or trait, connects indirectly to variability estimates. Similarly, Hansson, Juslin, & Winman (2008) correlated performance on WMC measures with estimated numeric confidence intervals of the income for fictitious companies. These researchers found that, while people provided interval estimates that were too small compared to the true intervals (i.e., underestimating variability), confidence intervals estimated by those with low WMC were significantly smaller than those with high WMC. Thus, two studies have clearly demonstrated a relationship between individual differences in working memory capacity with variability estimation performance.

We posit that some underlying cognitive ability is responsible for how people understand and respond to variability. Like Kareev, Arnon, and Horowitz-Ziegler (2002) suggested, this underlying ability could be working memory, since the number of "instances" in a distribution that one can hold in working memory will strongly influence understanding of the distribution as a whole. Even still, this study only addressed spatial stimuli, representing a part of how people encounter environmental variability. For example, numeracy, or a person's general mathematical ability, correlates with providing more accurate probability estimates for normal distributions (Rinne & Mazzocco, 2013). Even still, it is too early to probe exactly which underlying cognitive construct(s) may be responsible for the ability to understand and estimate variability, since no studies have examined the extent to which people show similar performance when making these variability judgements across different stimuli. The current study addresses this disparity.

CHAPTER 3 – PURPOSE AND RESEARCH QUESTIONS

Overall, prior research reveals that people often demonstrate sensitivity, albeit diminished, to different levels of variability. Further, responses to variability are typically less than the true amount, signifying underestimation. Researchers argue that factors such as anchoring and adjustment, sequencing of deviations from the distribution mean, and the amount of information sampled could all contribute to these trends. Cognitive factors implicated when estimating variability include how the information is encoded, the retention interval between encoding and response, cognitive biases, and the response methodology.

No studies have examined responses to variability across different types of stimuli in a repeated measures fashion. It is important for both theory and application to determine if, for the same individual, responses to variability are similar across stimuli. This empirical gap will provide evidence of the extent to which similar cognitive abilities may be responsible for a domain-general sense of variability, or if different cognitive abilities contribute to responses for specific stimuli. For example, if responses to numeric variability do not correlate with responses to spatial variability, then perhaps cognitive numeracy is responsible for the former, and spatial working memory for the latter.

The current experiment directly filled this gap by having participants estimate the variability and mean from multiple sets of continuous spatial and numeric instances. These responses were assessed for sensitivity, bias, and absolute error to identify whether the empirical trends presented above held in a within-participant design. To accomplish this, we used the SPUN spatial trajectory task and a new numeric variability estimation task called DigiVar, developed to be analogous to SPUN, but using continuously presented numeric stimuli.

Participants were recruited from Amazon Mechanical Turk (MTurk) to complete both tasks. During the SPUN task, we varied the speed and heading of the trajectories presented. Similarly, in DigiVar, we manipulated the mean and variability of number lists. To identify general, within-task trends, we examined average performance with changes in each manipulated variable. Subsequently, we correlated average performance across tasks to assess if there was a shared individual difference in the ability to estimate variability and means across the two modalities. Hypotheses central to the current study included:

- ***Hypothesis 1A.*** A significant between-task correlation will occur between the sensitivity of variability estimations[1];
- ***Hypothesis 1B.*** A significant between-task correlation will occur between variability bias (signed error) performance;
- ***Hypothesis 1C.*** A significant between-task correlation will occur between estimation error of the mean;
- ***Hypothesis 2.*** No significant within-task correlations will occur between variability sensitivity and estimation error of the mean; and
- ***Hypothesis 3.*** DigiVar task performance will reflect underestimation of variability.

[1] Initially this hypothesis was stated as so: "A significant between-task correlation will occur between the absolute error of variability estimations." Variability sensitivity, however, is a more robust way to assess understanding of variability across multiple levels compared to absolute error. In light of this, we have updated the hypothesis.

CHAPTER 4 – GENERAL METHODS

This experiment was done in two parts. The first set of data collected was considered a pilot study, in which 31 participants completed the SPUN task (Spatial Prediction with Uncertainty) and then the DigiVar (Digit Variability) task. Data from this pilot experiment were presented at the Human Factors and Ergonomics Society's 2018 Annual Conference (Spahr, Wickens, Clegg, Smith, & Williams, 2018). Given the results of the pilot experiment, the thesis committee suggested a follow-up experiment to counterbalance the two tasks, which would identify if any trends experienced a task order effect. See *Appendix A* for details about the thesis proposal.

Power Analysis

A power analysis was conducted to identify a proper sample size for the follow up experiment. This was based on the smallest meaningful effect size from the repeated measures ANOVA in the pilot study ($\eta_p^2 = 0.22$[2]), which corresponds to a large effect size. These cutoffs are based on the University of Cambridge's (2018) guidelines for the magnitude of multivariate eta squared effect sizes[3]. A power analysis at $a = .05$ and a power of 0.95 revealed that 16 participants were needed to achieve a large effect (Faul, Erdfelder, Buchner, & Lang [G*Power], 2009). The pilot study had 31 participants total; to properly counterbalance the two tasks, another 31 participants were recruited to complete the experiment with the tasks in the opposite order, making the total sample of 62 participants.

[2] This effect size occurred for the interaction of heading and speed on participants bias to variability (Wilk's $\Lambda = 0.78$, $F(1, 30) = 8.30$, $p = .01$, $\eta_p^2 = 0.22$).

[3] In a 2X2 repeated measures ANOVA, the multivariate eta squared equals the partial eta squared value $[Multivariate\ \eta^2 = 1 - \Lambda]$, where Λ is Wilk's Lambda (Horn, 2006).

Participants

All participants were recruited from Amazon Mechanical Turk (MTurk). For completing the study, they were compensated a base $6.00, with up to a $4.00 bonus for performance on specific SPUN trials. Participants were paid only if they completed both tasks; this was clearly stated in the instructions.

Analysis Plan

We first analyzed and presented the main effects of within-task manipulations using repeated measures ANOVAs. This was chosen for the ability to directly compare to other SPUN studies conducted by our lab (Herdener et al., 2016, 2017, 2018a, 2019a, 2019b). Task order was included in each ANOVA as a between-subjects variable.

To address each hypothesis, non-parametric tests were used when the variables failed to meet appropriate assumptions. For Hypotheses 1A, 1B, 1C, and 2, we utilized a Spearman's rho correlation technique. Hypothesis 3 was tested via a one-sample t-test. Analyses were conducted using SPSS (IBM, 2017) and R Studio (R Studio Team, 2015).

CHAPTER 5 – METHODS: SPUN

Participants completed a version of the SPUN task (Spatial Prediction with Uncertainty) with one practice block and four experimental blocks. Each experimental block consisted of two phases in which they encountered and predicted the mean and variability of spatial trajectories. These trajectories were built by a dot moving across the screen over three time points (See *Figure 3*).

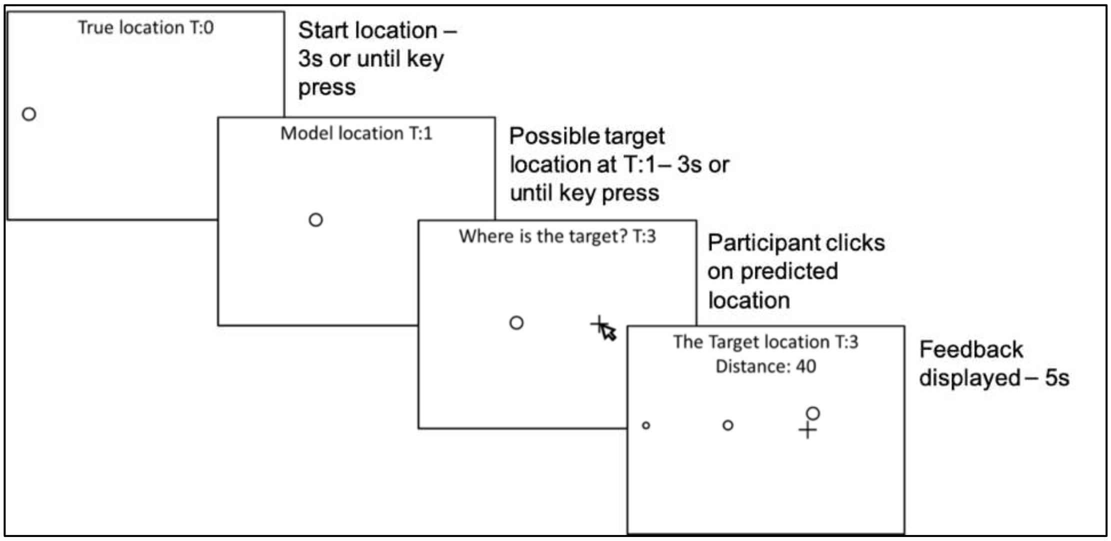
True location T:0
Start location – 3s or until key press
Model location T:1
Possible target location at T:1– 3s or until key press
Where is the target? T:3
Participant clicks on predicted location
The Target location T:3
Distance: 40
Feedback displayed – 5s

Figure 3. SPUN Trajectories. Trajectories for the SPUN paradigm are built by a location point moving across the screen over three time points. Participants predict the most likely location of T3.

Phase 1

Phase 1 (mean and variability prediction). During Phase 1 in each block, participants predicted the *mean* location of 20 trajectories. For each trajectory, participants first saw the initial position of a target, which we refer to as Time 0 [T0], and the position of that same target at a set duration later, referred to as Time 1 [T1]. These two positions represented the target moving from T0 to T1. Participants were then instructed to predict where the target would be at Time 3 [T3] by using the mouse to place the center of a circle at that point. This required respondents to extrapolate and make an implicit judgement where the target may be at Time 2 [T2], the location of which was not provided. After explicitly placing the circle where they believed the trajectory would be at T3, participants were instructed to adjust the diameter of that circle until it encompassed where they believed the target would be 75% of the time, given all trajectories that they had experienced up to that time within the block. This was an assessment of the *Phase 1 variability* of the trajectory endpoints in each block.

The 20 trajectories in each block were generated from the same underlying model, with a set speed (distance the dot could travel), and angle (potential for a heading change). When predicting the T3 location of the dot, a straight linear extrapolation of the first two points did not provide the correct answer. Instead, since those trajectories varied over trials, representing some degree of uncertainty, the best answer was one weighted by both the current trajectory points, and the centroid of the end of previously seen trajectories.

Phase 1 feedback. After estimating the mean T3 location and variability of possible T3 endpoints, participants received feedback on their performance. Specifically, they saw where the true T2 and T3 were for that single trajectory. This feedback was designed to help participants build a mental model of the trajectory variability behavior within each block, which they could then use to make more accurate predictions of both mean and variability during subsequent trials. After completing the set of 20 trajectories in Phase 1, participants moved onto Phase 2, which also asked for predictions of variability.

Phase 1 variables. Phase 1 assessed four key variables. First, for each single trial, we measured the distance between the actual T3 location and the participant's T3 prediction. This distance captured participants' *absolute spatial mean error*. To assess variability estimation performance, for each trial in each block, we measured how close participants' adjusted circle size was to the circle size that truly contained 75% of T3 locations. The adjusted circle size could be larger than the true size, calibrated, or smaller than the true size. For each block, we assessed *variability sensitivity* by subtracting the average estimate at low, from that at high variability. Finally, we distinguished between *Phase 1 variability bias* (the signed value of the difference; e.g., did participants overestimate or underestimate the circle size) and *absolute Phase 1 variability error* (e.g., overall, how accurate was the participants' given circle size).

Phase 2

Phase 2 (variability prediction). During Phase 2, participants again saw positions at T0 and then T1, followed by a circular, shaded area of fixed diameter and location, centered around the model T3 mean. Participants predicted the *probability* that the target was within that shaded area at T3. The underlying distribution did not change between Phase 1 and Phase 2, so the mental model that they constructed and received feedback on in Phase 1 applied for Phase 2 as well. Participants received no feedback during Phase 2 to avoid a confound of learning over both blocks. Participants estimated 20 probabilities during each Phase 2 block.

Phase 2 variables. Variables of interest in Phase 2 included variability sensitivity, bias, and absolute error. As with Phase 1, we obtained a Phase 2 *variability sensitivity* measure by subtracting average variability estimation (points in the circle) of large variability blocks from small variability blocks. The signed difference of [estimated probability – true probability] is an assessment of *variability estimation bias*. Positive values indicate that the participant overestimated variability, whereas negative values mean that the participant underestimated variability. Taking the absolute difference of this bias variable yields a measure of *absolute variability estimation error*; that is, how good participants were overall at estimating variability during Phase 2. No absolute error always implies no bias, however, low or zero bias does not necessarily imply no absolute error, as zero bias could result from the mean of both positive and negative bias trials. *Figure 4* presents an overview of both SPUN phases with key characteristics of each listed.

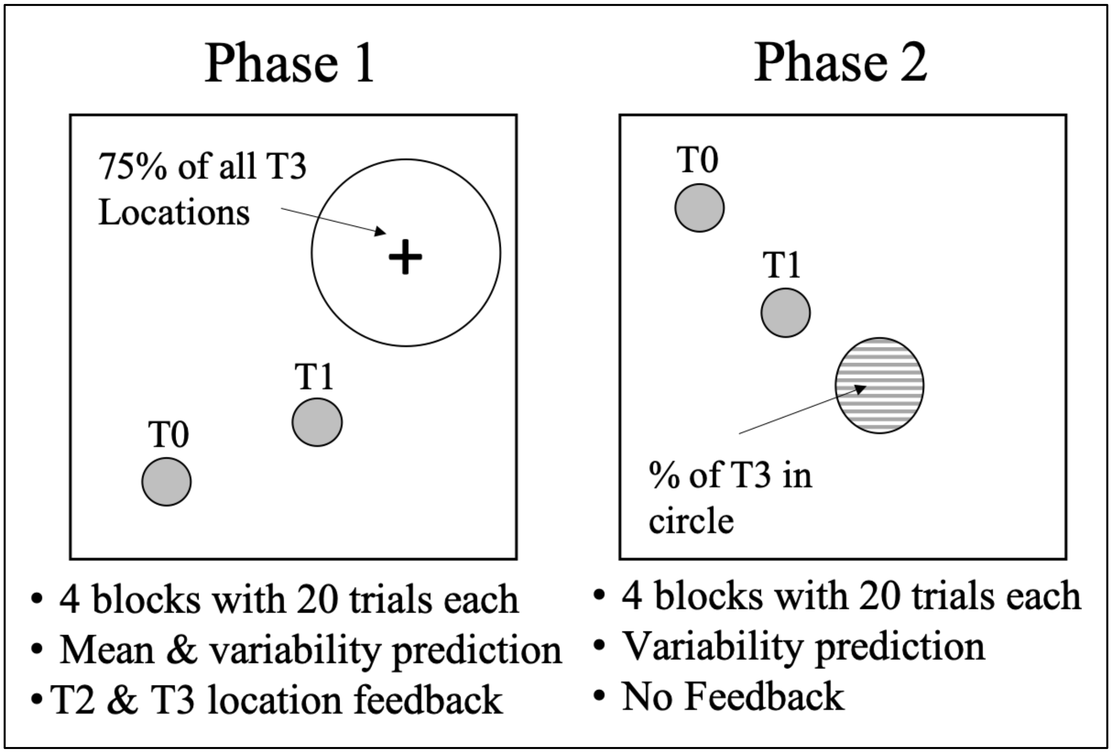
Phase 1
Phase 2
75% of all T3 Locations
T1
T0
T0
T1
% of T3 in circle
• 4 blocks with 20 trials each
• Mean & variability prediction
•T2 & T3 location feedback
• 4 blocks with 20 trials each
• Variability prediction
• No Feedback

Figure 4. Overview of the SPUN Task. The SPUN Task includes a Phase 1 and Phase 2. In Phase 1, participants predict the mean locations and variability of trajectories. In Phase 2, participants are shown a shaded circle and they predict the probability that the trajectory will be in that shaded area at T3.

Financial incentives. Based on circle adjustment performance in Phase 1, participants were able to earn up to $4.00 of financial compensation across the four SPUN blocks. Specifically, participants were informed that on the last four trials of each block, the closer participants adjusted to the true 70% variability of T3 points, the more bonus compensation they received. This bonus system was employed to encourage attention to the task.

SPUN Trajectories

Trajectories for both phases were drawn from an underlying model with a normally distributed angle and speed. We manipulated the parameters in each Phase1 - Phase2 trajectory model such that participants experienced four total 2-way combinations of high or low angle variability and high or low speed variability. These conditions naturally contributed to differences in the dispersion of T3 points between blocks (see *Figure 5*). The four blocks were presented in random order, and the conditions are henceforth labeled:

- HA HS (high angle variability, high speed variability),
- HA LS (high angle variability, low speed variability),
- LA HS (low angle variability, high speed variability), and
- LA LS (low angle variability, low speed variability).

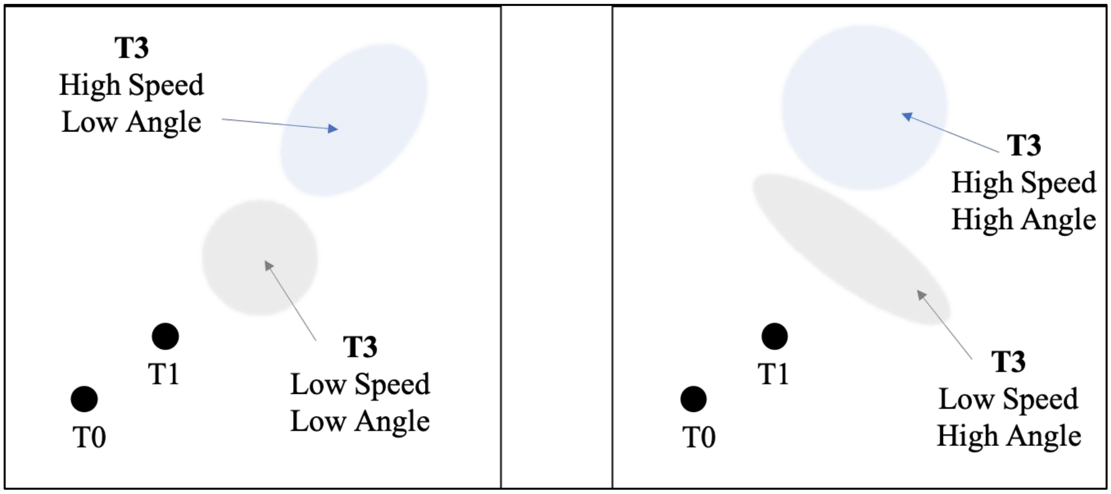

Figure 5. SPUN Distributions. The shaded areas exemplify different distributions of T3 locations given combinations of angle and speed.

SPUN Methods Summary

In each of 4 Phase 1 - Phase 2 blocks, participants viewed the beginning of 20 semi-random spatial trajectories, wherein a dot moved from T0 to T1. In Phase 1, participants estimated the average location of the dot at T3 for each trial, and then estimated the 75% variability in T3 locations by sizing a circle around their provided average location. In Phase 2, the location of the dot was presented at T0 and T1, plus a shaded circle of fixed diameter. Participants estimated the likelihood of the dot falling in that circle at T3, again assessing trajectory variability. As described earlier, there was a between-participant manipulation of task order. If applicable, this will be discussed at the end of the presentation of each dependent variable.

CHAPTER 6 - RESULTS: SPUN

Variability Estimation

One of the main outcomes from this experiment was estimates of spatial variability. Regarding these estimations, we derived measures of sensitivity, bias, and absolute error during SPUN Phase 1 and Phase 2. In all graphs, the intervals around each point represent the standard error of the mean.

Distribution of responses: Circle size. For Phase 1, participants responded to spatial variability by sizing a circle to capture 75% of all trajectory endpoints. *Figure 6* presents the distribution of average circle sizes across all conditions. The histogram reveals that most frequently, average circle sizes ranged between 80 pixels and 110 pixels, with no apparent outliers.

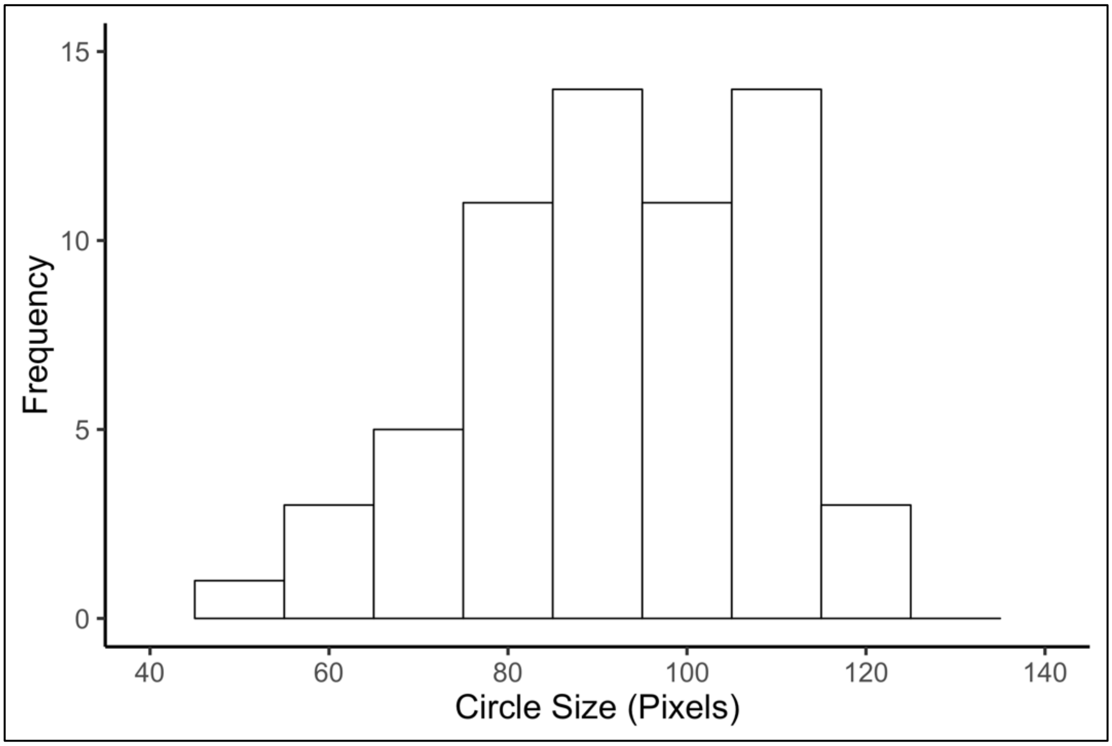

Figure 6. Histogram of Circle Sizes (Pixels). This histogram shows participants' average given circle size across all conditions.

Phase 1 variability sensitivity: Circle size. Examining the size of circles for each condition identifies the extent to which participants were sensitive to spatial variability in Phase 1. If average circle sizes at all increased as the amount of true variability increased, we can say that participants were at least somewhat sensitive to variability. *Figure 7* presents both the true and participant estimated average adjusted circle size as a function of speed and angle variability. The true circle sizes are represented by the thicker lines, and the participant estimated circle sizes are represented by the thinner two lines.

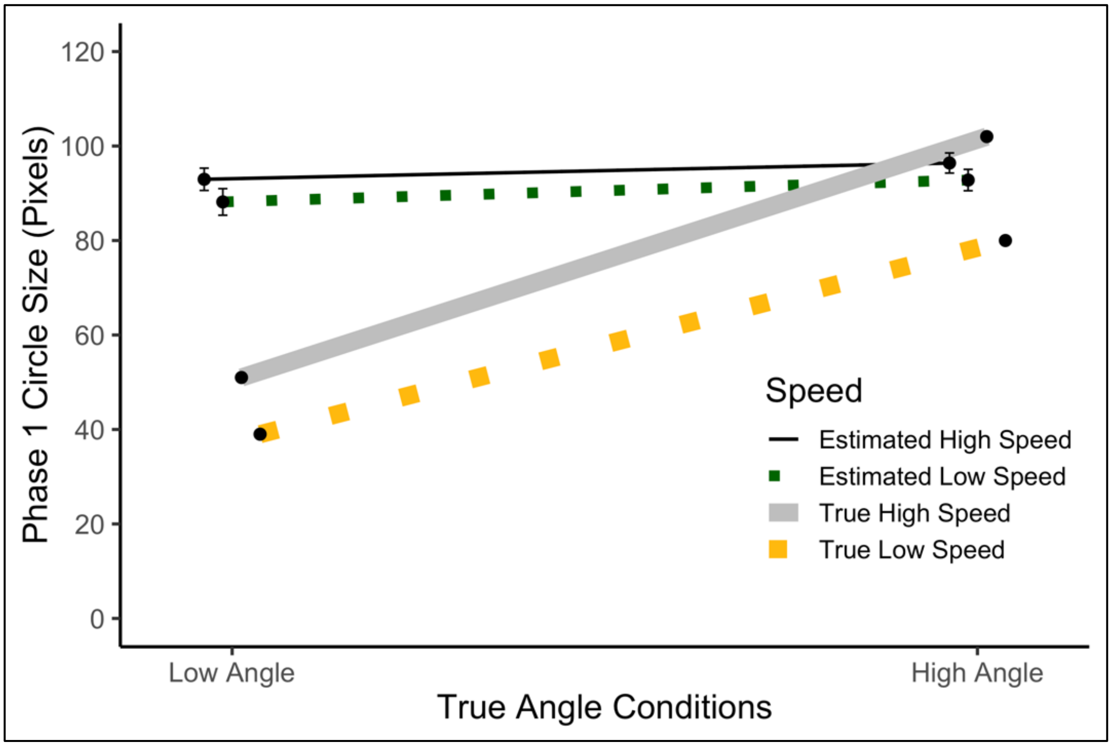

Figure 7. Phase 1 Variability Sensitivity [Average Adjusted Circle Size]. On average, the circle sizes increased as true variability increased, suggesting that participants were somewhat sensitive to true variability.

Figure 7 shows that with increasing true angle variability, participants gave slightly larger circle size responses (exemplified by the positive slope of the two lines), meaning that they were somewhat sensitive to increasing angle variability. Similarly, the difference between the low- and high-speed lines suggests that participants were also somewhat sensitive to increasing speed variability.

A 2X2 MANOVA supported these observations, with both a significant main effect of angle variability [Wilk's Λ = 0.90, F(1, 60) = 7.02, $p = .01$, $\eta_p^2 = 0.11$] and speed variability [Wilk's Λ = 0.84, F(1, 60) = 11.72, $p < .01$, $\eta_p^2 = 0.16$] on participant's average adjusted circle size. The interaction between speed and angle was not significant [Wilk's Λ = 0.99, F(1, 60) = 0.23, $p = .64$, $\eta_p^2 = 0.00$]. There were no significant order effects noted for average adjusted circle size.

Degree of sensitivity. We quantified the amount of sensitivity shown in Phase 1 by creating a ratio of estimated to true variability using the difference in the lowest variability condition (low-speed, low-angle) to the highest variability condition (high-speed, high-angle). The average estimated circle size was 88.17 pixels for low variability and 96.17 pixels for high variability, for a difference of 8.24 pixels. The true circle size was 39 pixels for low variability and 102 pixels for high variability, with a difference of 63 pixels. Consequently, the difference ratio of estimated to true variability for Phase 1 was $Sensitivity_{P1} = \frac{8.24}{63}$.

Phase 1 variability bias. We calculated variability bias by subtracting the true circle size for each condition from the average Phase 1 estimated circle size. Positive values signify average overestimation of variability, and negative values signify average underestimation of variability. Zero corresponds to no bias. *Figure 8* presents the average bias to variability for each condition.

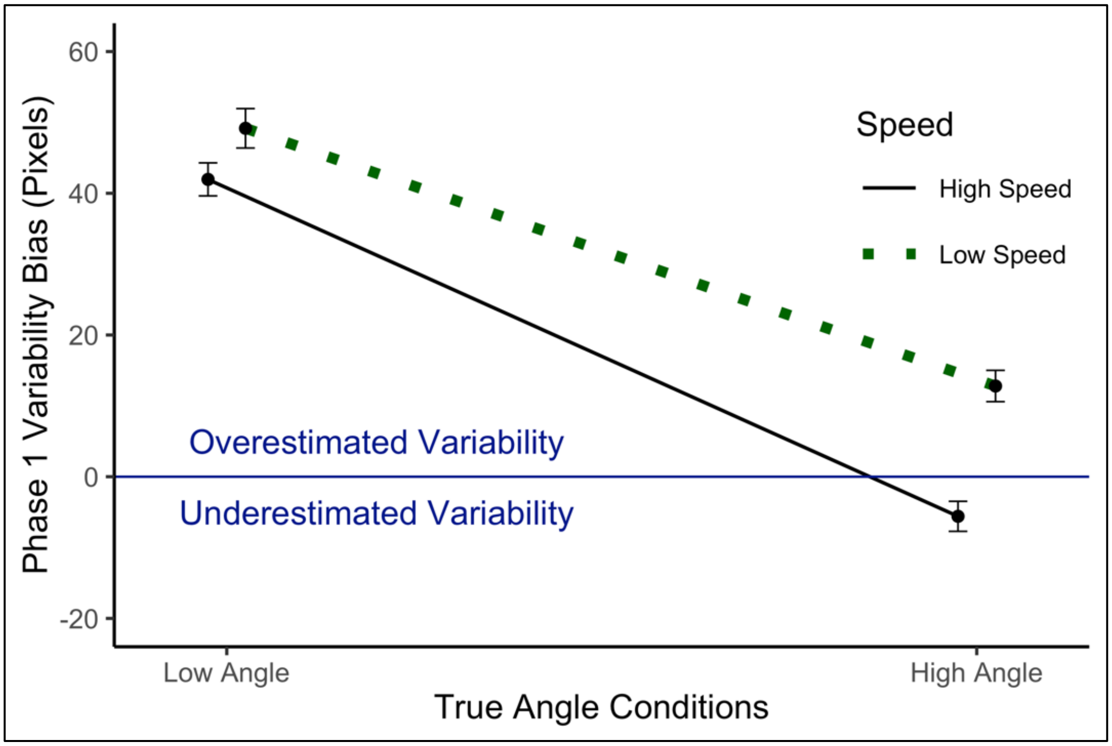

Figure 8. Phase 1 Average Variability Bias. On average, participants' responses became less biased as true variability increased.

Figure 8 shows that participants on average, overestimated spatial variability on three out of four conditions. For the high-speed and high-angle condition, participants slightly underestimated variability. The negative slope of the lines suggests that estimates were less biased as variability increased.

A 2X2 MANOVA revealed a significant, large, main effect of true angle variability [Wilk's Λ = 0.07, F(1, 60) = 756.54, p < .01, η_p^2 = 0.93] and a large, significant, main effect of true speed variability [Wilk's Λ = 0.36, F(1, 60) = 108.70, p = .01, η_p^2 = 0.64] on average participant Phase 1 variability bias. The interaction between speed and angle variability was also significant [Wilk's Λ = 0.75, F(1, 60) = 19.75, p < .01, η_p^2 = 0.25]. No significant order effects occurred for this variable.

Absolute Phase 1 variability error. Absolute variability error was calculated by taking the absolute value of Phase 1 bias. Smaller values correspond to *less* overall error, or *more* *precise* responses to variability. *Figure 9* presents participant's average absolute variability error for each of the four conditions. The negative slope means that, for both low-speed and high-speed variability, as angle variability increased, average absolute error decreased.

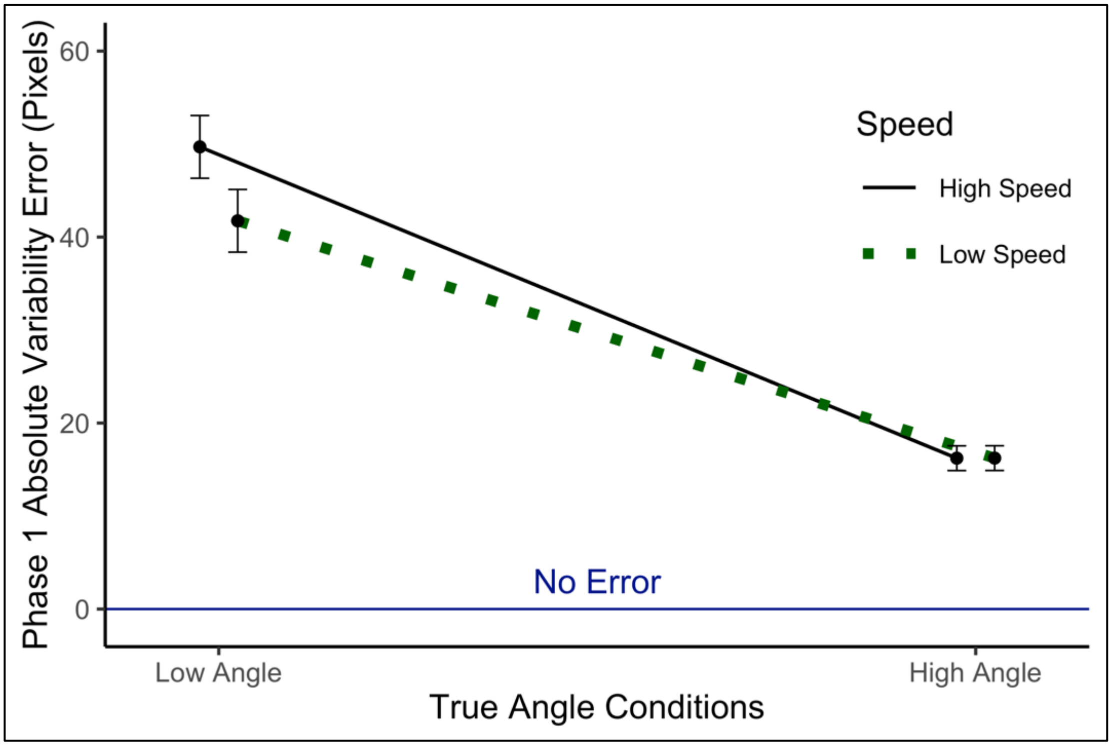

Figure 9. Absolute Phase 1 Variability Error. For both low-speed and high-speed variability, as true angle variability increased, participants' average absolute error decreased.

A 2X2 MANOVA revealed a significant main effect of true angle variability [Wilk's Λ = 0.29, F(1, 60) = 145.55, *p* < .01, η_p^2 = 0.71] and large, significant, main effect of true speed variability [Wilk's Λ = 0.80, F(1, 60) = 14.93, *p* < .01, η_p^2 = 0.20] on average absolute error. The interaction between speed and angle variability was not significant [Wilk's Λ = 0.96, F(1, 60) = 0.87, *p* = .35, η_p^2 = 0.01]. No significant order effects occurred for this variable.

Distribution of responses: Estimated variability. For Phase 2, participants responded to spatial variability by providing a percentage of T3 points they believed would fall in a circle of fixed diameter. *Figure 10* presents the distribution of estimated percentages, averaged across all conditions. The histogram reveals that most frequently, average estimates ranged between 0.1 and 0.5.

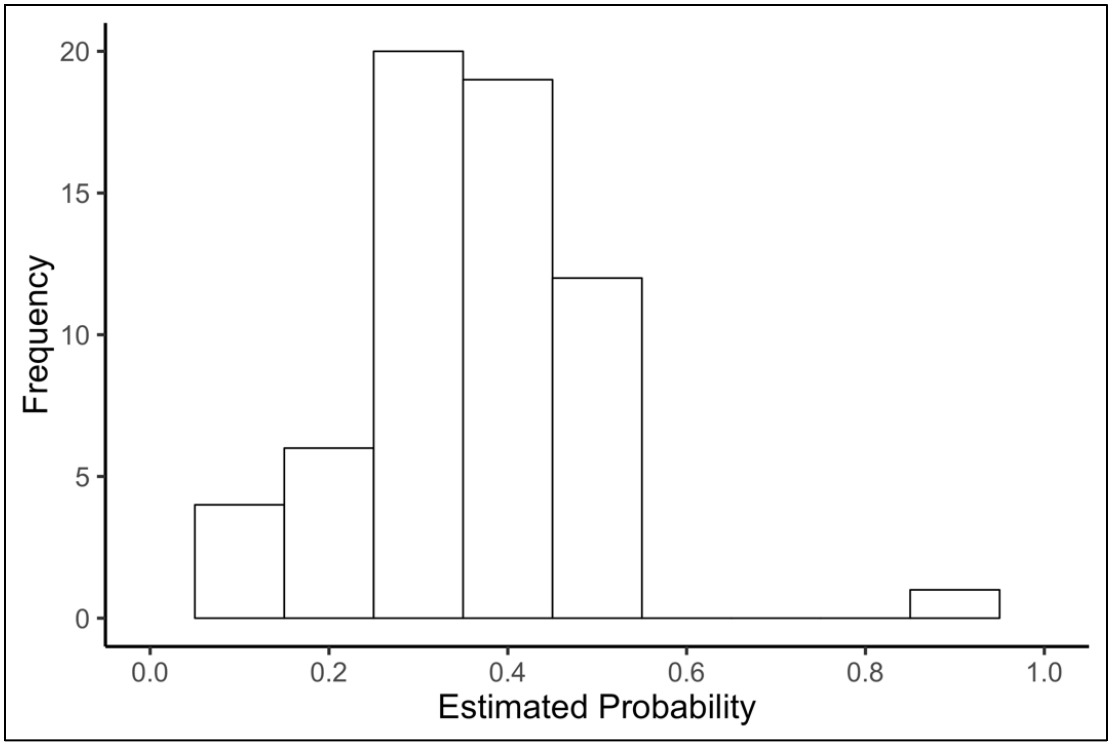

Figure 10. Histogram of Probability Estimates. In this histogram, responses for each participant were averaged across conditions. Most frequently, these estimates ranged from 0.1 to 0.5.

Phase 2 variability sensitivity: Estimated percentage within the circle. Using the estimated probability variable, we derived the Phase 2 equivalent of variability sensitivity. *Figure 11* presents both the true and average estimated T3 probability for each condition, and suggests that, as angle variability increased, participants' average probability estimates decreased. As such, participant's responses were slightly sensitive to changes in true angle variability.

A 2X2 MANOVA revealed a significant main effect of true angle variability [Wilk's Λ = 0.58, F(1, 60) = 43.24, *p* < .01, η_p^2 = 0.71] on average variability assessments. The main effect of true speed variability [Wilk's Λ = 0.96, F(1, 60) = 2.54, *p* = .12, η_p^2 = 0.04] and the interaction between speed and angle variability [Wilk's Λ = 0.99, F(1, 60) = 0.33, p = .57, η_p^2 = 0.01] were both not significant.

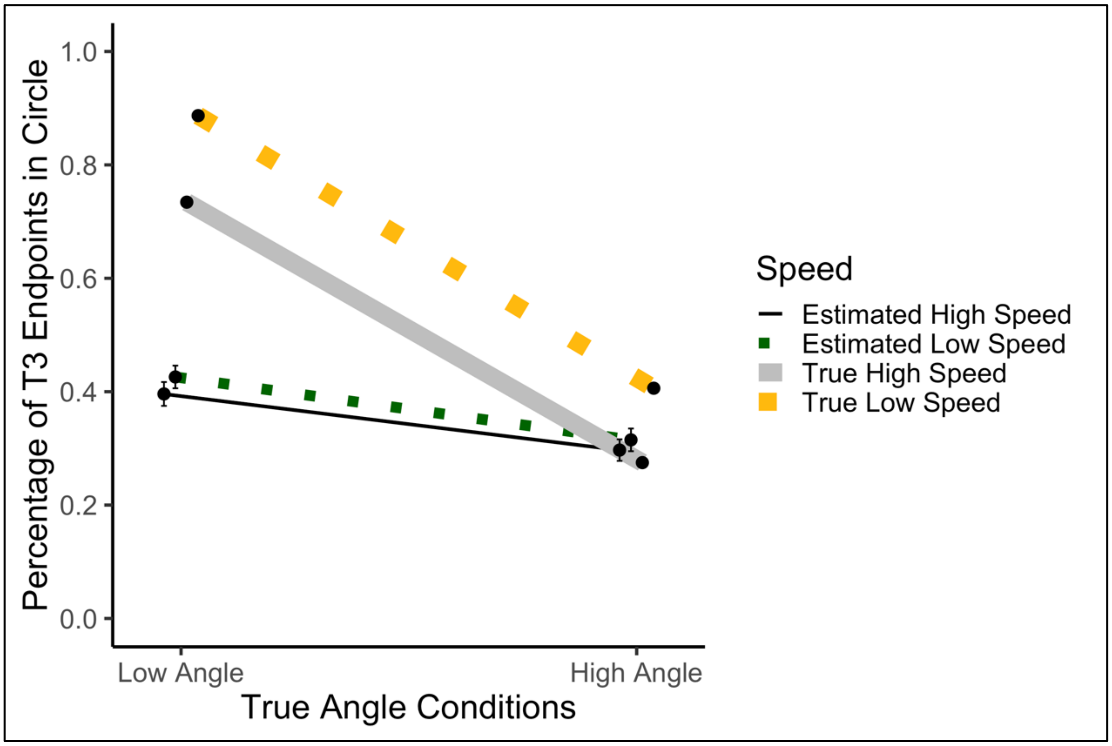

Figure 11. Phase 2 Sensitivity to Variability [Estimated Percentage in Circle]. As angle variability increased, participants' average probability estimates decreased. This suggests that participants were slightly sensitive to changes in true angle variability.

Phase 2 variability assessment and task order. The only significant interaction with task order was again with the speed manipulations [Wilk's Λ = 0.93, F(1, 60) = 4.67, p = .04, η_p^2 = 0.07]. Because speed did not have a main effect on responses, we did not interpret this interaction.

Degree of sensitivity. We again created a ratio of estimated to true variability to quantify the amount of sensitivity shown in Phase 2. The average estimated probabilities in Phase 2 were 0.43 for low variability and 0.30 for high variability, leading to a probability difference of 0.13. The true probability values were 0.89 for low variability and 0.27 for high variability, with a difference of 0.62. The difference ratio of estimated to true for Phase 2 was $Sensitivity_{P2} = \frac{0.13}{0.62}$.

Phase 2 variability bias. *Figure 12* presents the Phase 2 average bias across conditions. Positive values signify average overestimation of variability, negative values signify average underestimation of variability, and zero corresponds to no bias.

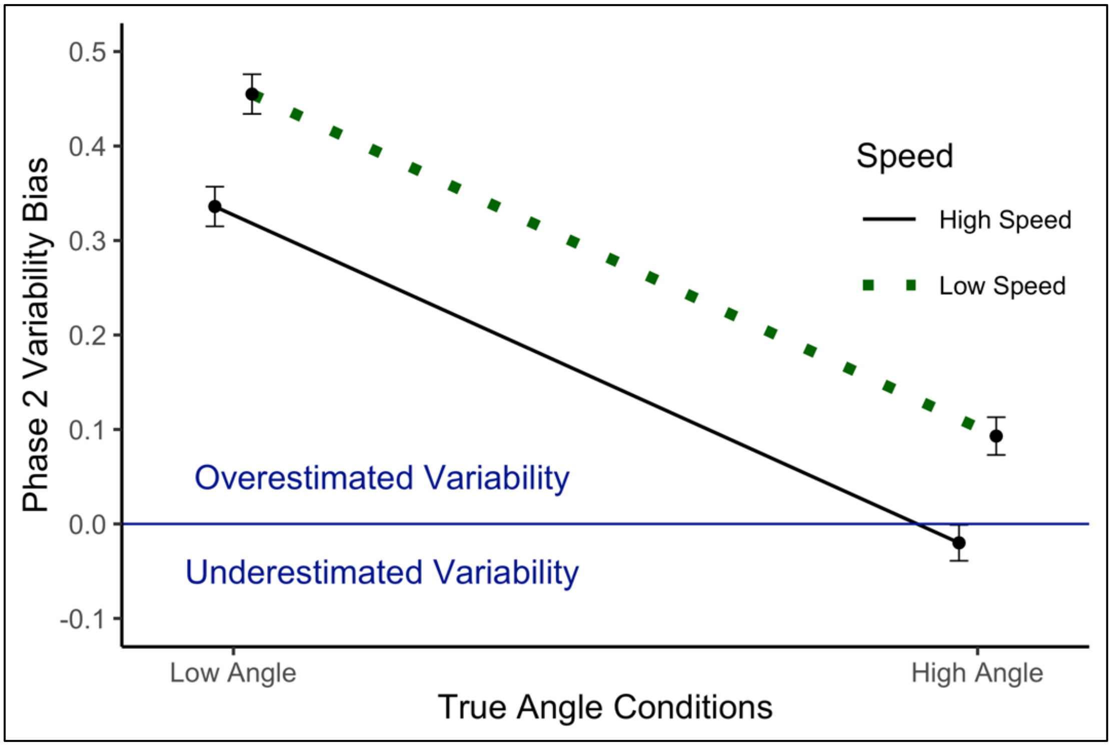

Figure 12. Phase 2 Average Variability Bias. On average, as true speed and angle variability increased, responses became less biased.

Figure 12 suggests that, as with the Phase 1 variability estimations, in three out of four conditions, on average, participants overestimated variability. For the high-speed, high-angle condition, participants were very nearly calibrated, but slightly underestimated variability. A 2X2 MANOVA revealed a significant main effect of true speed variability [Wilk's Λ = 0.36, F(1, 60) = 107.56, $p < .01$, $\eta_p^2 = 0.64$] and true angle variability [Wilk's Λ = 0.13, F(1, 60) = 391.87,

$p < .01$, $\eta_p^2 = 0.87$] on participant Phase 2 variability bias. The interaction between speed and angle variability was not significant [Wilk's Λ = 0.99, F(1, 60) = 0.08, $p = .77$, $\eta_p^2 = 0.00$].

Phase 2 variability bias and task order. The only significant interaction with task order was again with the true speed conditions [Wilk's Λ = 0.90, F(1, 60) = 6.93, $p = .01$, $\eta_p^2 = 0.10$]. Collapsed across angle conditions, participants that completed the SPUN task first had an average overestimation bias of 0.31 (SD = 0.16) for the low-speed condition and an overestimation bias of 0.16 (SD = 0.15) for the high-speed condition. Conversely, the participants who completed the DigiVar task first had an average probability of 0.24 (SD = 0.17) for the low-speed condition and an average probability of 0.15 (SD = 0.16) for the high-speed condition. The difference between the low-speed and high-speed condition for participants who completed the SPUN task first was 0.15, whereas the difference between speed conditions for participants who completed the SPUN task second was 0.09.

Absolute Phase 2 variability error. Similar to Phase 1, we determined absolute Phase 2 variability error by taking the absolute value of bias. *Figure 13* presents the absolute error for all four conditions. For both speed conditions, average absolute error decreased as true angle variability increased. A 2X2 MANOVA suggested a significant main effect of true angle variability [Wilk's Λ = 0.28, F(1, 60) = 158.35, $p < .01$, $\eta_p^2 = 0.73$] and true speed variability [Wilk's Λ = 0.52, F(1, 60) = 55.06, $p < .01$, $\eta_p^2 = 0.48$]. Further, the interaction between speed and angle variability was also significant [Wilk's Λ = 0.83, F(1, 60) = 12.27, $p < .01$, $\eta_p^2 = 0.17$]. No significant order effects occurred for this variable.

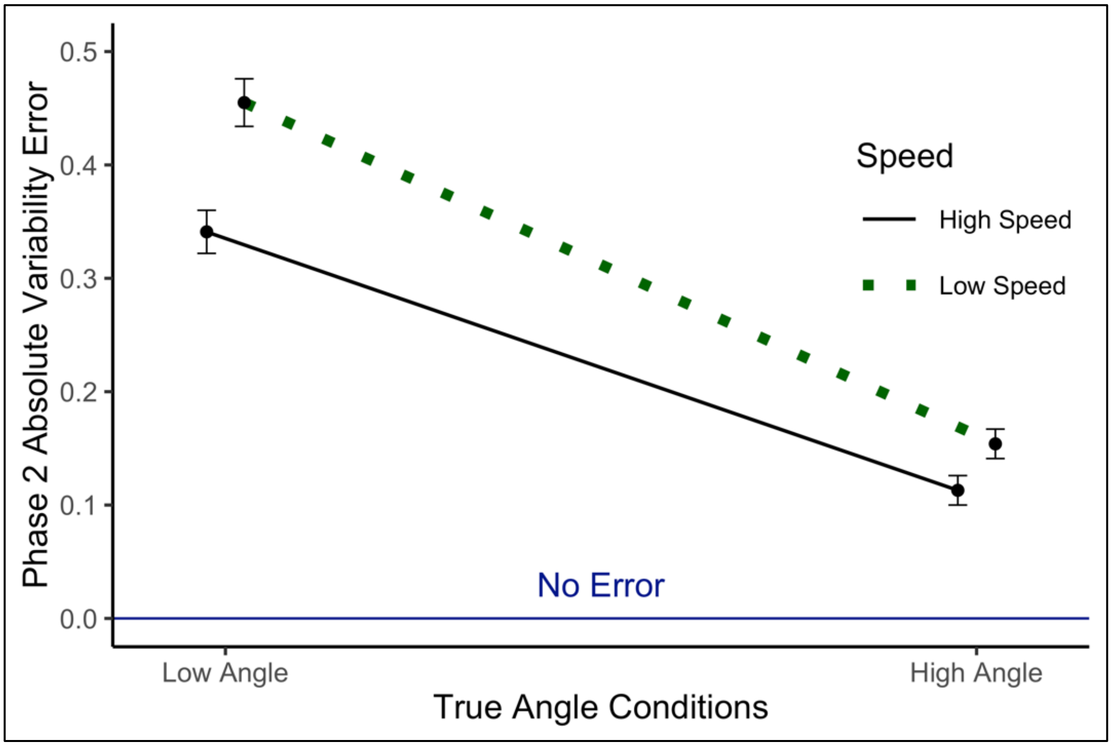

Figure 13. Absolute Phase 2 Variability Error. On average, as true angle and speed variability increased, participants' absolute response error decreased.

Mean Estimation

Distribution of responses: Mean prediction error. During Phase 1, participants predicted the trajectory mean locations. Mean prediction error was computed by taking the absolute distance between the estimated and true mean location. *Figure 14* presents the distribution of mean error averaged across all conditions. The histogram reveals that most frequently, average estimates ranged between 50 pixels and 110 pixels.

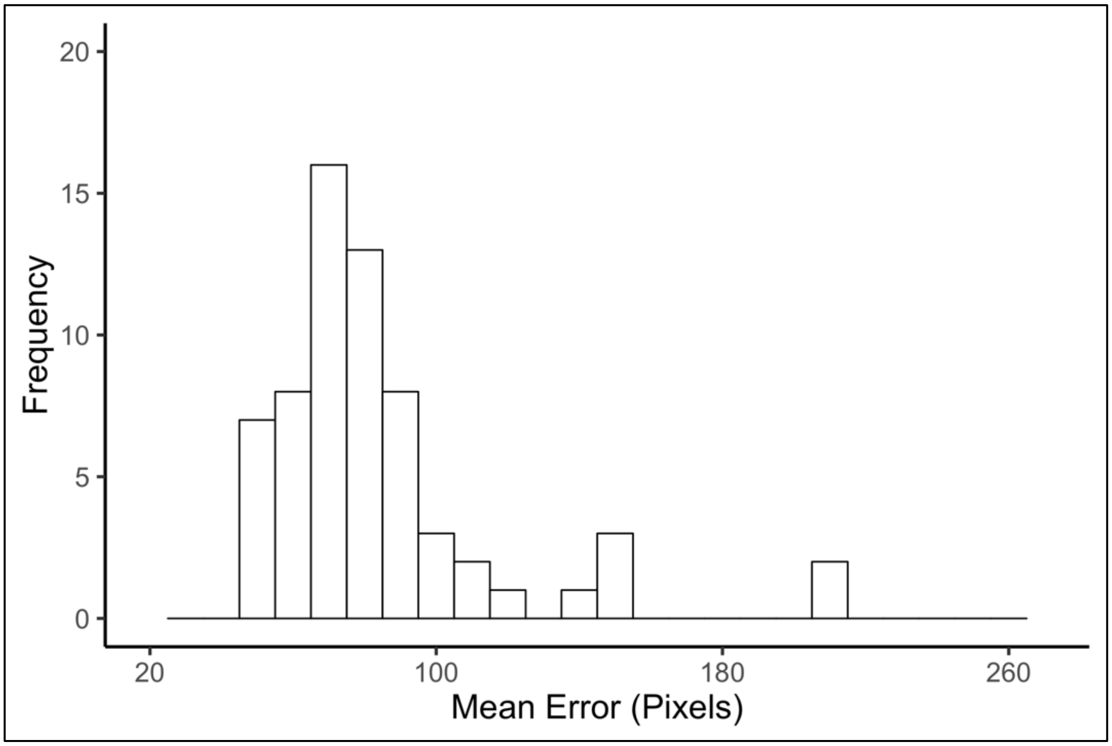

Figure 14. Histogram of Spatial Mean Error. In this histogram, responses from each participant were averaged across conditions. Most frequently, these estimates ranged from 0.1 to 0.5.

Spatial mean prediction error. On each trial participants predicted the T3 mean location. Mean prediction error was computed by taking the absolute distance between the estimated and true mean location. Larger values correspond to more error. Perfect mean precision (i.e., 0 error) is impossible in this task because of underlying variability in the trajectories. *Figure 15* presents the average spatial mean absolute prediction error across the four conditions.

Average spatial mean prediction error increased as true angle variability increased. Similarly, the difference between the low speed (i.e., dashed green line) and high speed (i.e., solid black line) conditions suggest that spatial prediction error increased as speed variability increased.

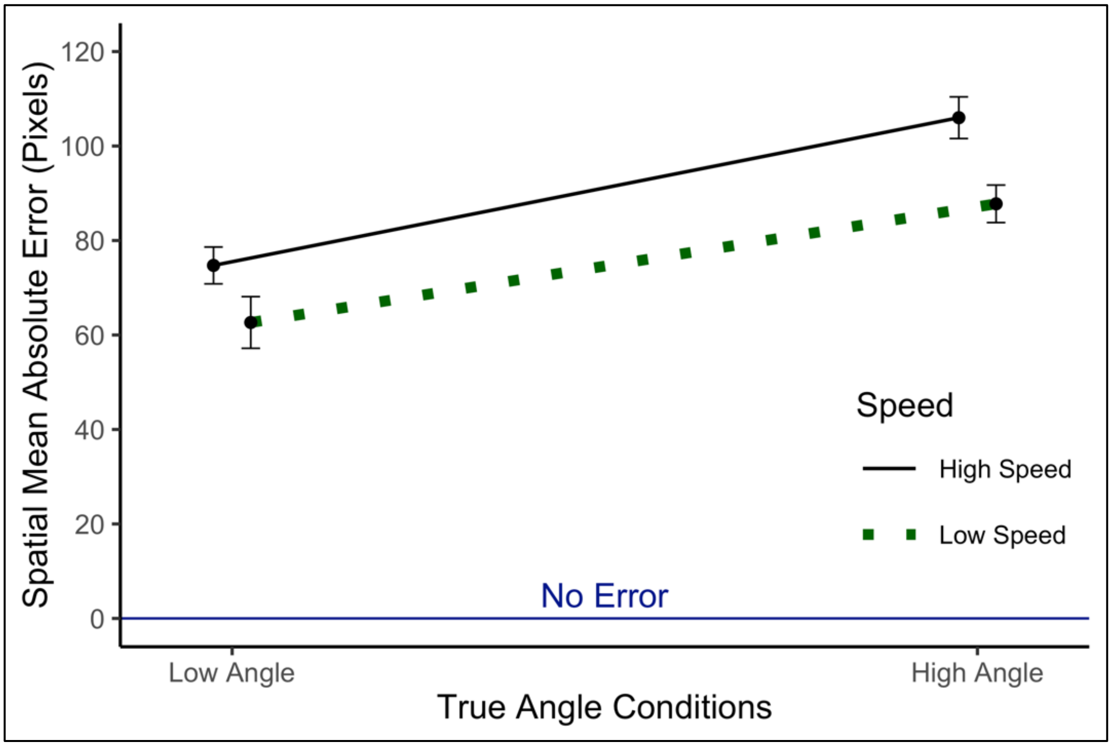

Figure 15. Absolute Spatial Mean Error. On average, participants' spatial mean prediction error increased as true angle and true speed variability increased.

A 2X2 MANOVA revealed both a significant main effect of true angle variability [Wilk's $\Lambda = 0.44$, F(1, 60) = 77.02, $p < .01$, $\eta_p^2 = 0.56$] and true speed variability [Wilk's $\Lambda = 0.70$, F(1, 60) = 26.16, $p < .01$, $\eta_p^2 = 0.30$] on participant's spatial mean prediction error. The interaction between speed and angle was not significant, exemplified by the relatively parallel low- and high-speed condition lines [Wilk's $\Lambda = 0.98$, F(1, 60) = 1.23, $p = .26$, $\eta_p^2 = 0.02$].

Mean prediction and task order. The only significant interaction with task order was with the true speed conditions [Wilk's $\Lambda = 0.89$, F(1, 60) = 7.45, $p = .01$, $\eta_p^2 = 0.11$]. Collapsed across angle conditions, participants who completed the SPUN task first had a spatial mean prediction error of 77.49 (SD = 42.74) for the low speed condition and a spatial mean prediction error of 84.55 (SD = 30.72) for the high-speed condition. Conversely, the participants who completed the DigiVar task first had a spatial mean prediction error of 72.93 (SD = 28.75) for the low speed condition and a spatial mean prediction error of 96.17 (SD = 34.55) for the high-speed condition.

Since the lower speed conditions led to less spatial mean prediction error, we computed a “low speed advantage” by subtracting the error for the low speed condition from the error at high speed. When SPUN was first, the low speed advantage was 7.06 pixels and when SPUN was second, the low speed advantage was 23.24 pixels. This suggest that there was less of a difference between the low- and high-speed conditions when participants completed SPUN first.

CHAPTER 7 - DISCUSSION: SPUN

Results from the SPUN task revealed a number of empirical trends regarding how participants estimated spatial variability and means.

Task Order

While some interactions occurred between task order and true speed variability, in these cases, the main effect of speed itself was relatively small or entirely non-significant (i.e., Phase 2 variability sensitivity, mean prediction error). These trends imply that completing SPUN first versus second had only a small effect on overall performance.

Sensitivity to Spatial Variability

The difference ratio of estimated to true variability for Phase 1 was $Sensitivity_{P1} = \frac{8.24}{63}$. This suggests that participants were sensitive to changing variability, as seen through the positive denominator, but the amount of sensitivity was much lower than calibration (8.24 compared to 63). For Phase 2 the sensitivity ratio was $Sensitivity_{P2} = \frac{0.13}{0.62}$. 0.21Similar to the interpretation for Phase 1, this ratio demonstrates that participants were sensitive to changing variability, but again, much lower than calibration (0.13 compared to 0.62).

Figure 16 below presents a conceptual diagram of the sensitivity demonstrated in both Phase 1 (gold) and Phase 2 (green) with the black diagonal line representing perfect calibration (estimated variability = true variability). This figure demonstrates that in both phases, participants did change their responses to variability appropriately, seen through the positive slopes of the Phase 1 and Phase 2 lines, but responses were under-sensitive (both had a flatter slope compared to the black line). Participants demonstrated better sensitivity to variability in Phase 2 compared to Phase 1.

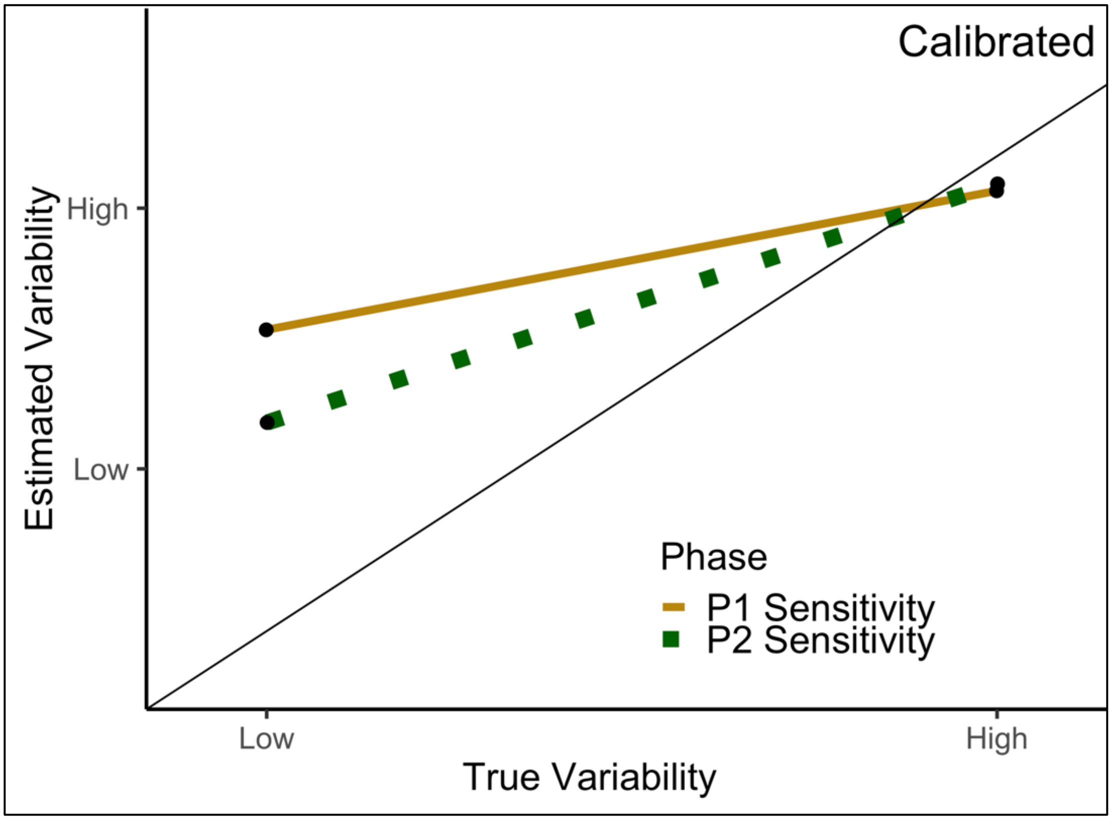

Figure 16. SPUN Sensitivity Diagram. This diagram shows how estimated variability changed as a function of true variability for both Phase 1 and Phase 2. As true variability increased, estimations of variability became more calibrated.

Speed and angle sensitivity. Additional sensitivity trends can be gleaned from the effect sizes of angle and speed on Phase 1 and Phase 2 sensitivity measurements. Specifically, the partial eta-squared effect sizes for speed and angle include the following:

- Sensitivity to angle variability: $\eta_p^2 = 0.11$ (Phase 1) and $\eta_p^2 = 0.71$ (Phase 2)
- Sensitivity to speed variability: $\eta_p^2 = 0.16$ (Phase 1) and $\eta_p^2 = 0.04$ (Phase 2).

During Phase 1, there was a larger effect size for speed variability compared to angle variability ($\eta_p^2 = 0.16$ versus 0.11), suggesting that participants responded with more sensitivity to speed changes. This trend was reversed in Phase 2, with a much larger effect size for angle compared to speed variability ($\eta_p^2 = 0.71$ versus 0.04). Such results suggest a dissociation between the sensitivity when responding via an interval estimation as in Phase 1, and making a probability estimate as in Phase 2. The MOVE model indeed incorporates response methodology as a key component of variability estimation, with sizing an interval relying on perceptual abilities and providing a probability judgement relying more on memory systems. Hypothetically, it was easier for participants to perceptually notice changes in speed variability compared to angle variability, because speed changes vary on only one dimension, distance, whereas angle changes involve both distance and direction changes. Further, The Phase 1 response is fixed as a circle, with the only changes possible including increasing or decreasing the diameter. As such, it is easier to incorporate changes in the distribution of T3 locations in terms of speed/distance, but harder in terms of angle, since the circle shape cannot change.

In Phase 2, changes in true angle variability led to a much larger effect size, suggesting that the translation from a perceptual to a memory-based response system made angle changes much more salient to participants. Angle changes can be more easily visualized compared to speed changes, as the latter requires more mental calculation (Herdener, et al., 2016). Given that the probability estimates required in Phase 2 already forced mental operations for any response, participants could have simply been left more sensitive to angle changes when making their estimates.

Variability Bias

Another important aspect of variability estimation is the bias expressed by participants, that is, whether they over- or underestimated variability. For both Phase 1 and Phase 2, participants overestimated variability in all conditions except high-angle and high-speed. Overestimating variability corresponds to participants responding as if the dispersion of T3 points is larger than it really is; in Phase 1, this is expressed as sizing a *larger* circle than calibrated, and in Phase 2, this is expressed as estimating a *smaller* probability of T3 falling in the circle than optimal. To aid discussion of the variability bias expressed in both phases, we calculated the average percent variability bias (overestimated variability [OEV] and underestimated variability [UEV]) for all conditions based on the true values, presented in *Table 1*.

Table 1. Variability Bias Across Phases

Condition	Phase 1	Phase 2
Low-Speed, Low-Angle	126% OEV	108% OEV
Low-Speed, High-Angle	16% OEV	27% OEV
High-Speed, Low-Angle	82% OEV	85% OEV
High-Speed, High-Angle	5.5% UEV	7.4% UEV
Overall Average	54.63% OEV	53.15% OEV

Across all conditions, participants overestimated variability by 54.63% on Phase 1 and 53.15% in Phase 2. As true variability increased, the magnitude of overestimation decreased. In the highest variability condition, participants, on average, slightly underestimated variability. These findings are completely opposite our previous SPUN studies (Herdener et al., 2016, 2017, 2018, 2018a, 2019b; Pugh et al., 2018) which each found clear underestimation of variability.

The trends of variance overestimation represent a failure to replicate not only past SPUN experiments, but also other non-numeric studies discussed above which noted a general underestimation of variability (Hofstatter, 1939; Lathrop, 1967). This observation could have a number of explanations. First, the SPUN task used here differed in some key ways from the task as used in previous studies. In Herdener et al. (2016), during Phase 1, participants only placed a circle at the estimated T3 location, without making a variability judgement. During the current SPUN paradigm, we ask participants to assess the mean location ***and*** variability concurrently (i.e., by positioning ***and*** sizing the circle at the same time). This additional variability judgement not only provided additional practice making such assessments, which may have influenced the shift in results compared to Herdener et al. (2016). Furthermore, making variability judgement across both phases creates a stronger, ongoing memory of variability than was the case in the prior paradigm in which only Phase 2 variability was assessed (i.e., after all instances had been encountered). This ongoing memory of variability, reinforced by continuous responses, best explains why the Phase 1 and Phase 2 variability results were so similar, with clear overestimation of variability for all conditions except high-speed, high-angle, during which participants provided almost calibrated responses.

Memory encoding theory, or the process of creating memories, can partially explain these results (Wickens et al., 2013). While participants in the current study and in Herdener et al. (2016) were exposed to the same amount of variability through Phase 1 and Phase 2, in the latter, they were only required to respond to that variability during Phase 1. Conversely, responding to variability throughout each phase of the task, as required in the present study, may have encouraged participants to differentially consider and encode patterns of variability, leading to the reversal of results seen here. It is important to note that this increase in the repetition of responding to variability didn’t make participants remember that behavior more accurately than that in Herdener et al. (2016). Indeed, as repetition alone doesn’t improve memory (Roediger, 2008), it makes sense that we didn’t see participants becoming less biased with more chances to response. Instead, they simply changed the directionality of bias to be one of over- rather than under-estimation.

An additional difference between this study and others we have run focuses on financial incentive. In this study, participants could earn a bonus for properly calibrated responses during the last four trials of Phase 1, a fact stressed during instructions and on the screen for each possible bonus trial. Such incentives may have reinforced an “inclusive” cognitive strategy to absolutely make sure and capture 70% of the T3 endpoints. Such a strategy would force an overestimation of variability to at least hit 70%, but often overshooting this value, particularly when the T3 endpoints were less dispersed (Low-Speed and Low-Angle conditions). While this specifically concerned Phase 1, such a strategy could have an echo effect through to Phase 2, leading to the same outcome phenomenon.

Estimating Spatial Means

In the current study, participants estimated the mean T3 trajectory location by placing the center of a circle on the most likely location for T3. Performance for this prediction was expressed through the distance between the predicted and actual T3 location. Overall, prediction error increased as true variability (both speed and angle) increased. This suggests that participants, on average, were better at predicting the mean trajectory when variability was lower. Across all conditions, average mean prediction error was 82.79 pixels (SE = 3.65).

Performance on mean estimation in the present study are similar to those from past SPUN studies using an analogous methodology. In Herdener et al. (2016), average mean error was 96.02 pixels (SE = 1.71) collapsed across all levels of heading and speed variability; authors noted that this average performance was still significantly higher than the optimal performance of an average error of 73 pixels. As noted in the results above, participants in the present study made significantly better mean predictions compared to Herdener et al. (2016), but the average mean error was still significantly greater than the optimal 73 pixels.

The major difference between the SPUN methodology used here and the one in Herdener et al. (2016) was the addition of the Phase 1 variability measure (i.e., having participants size their circle in Phase 1 to capture 70% of the T3 endpoints). While this was indeed a change in Phase 1 that could have influenced overall performance, it seems unlikely that an additional step that had participants focus on variability would have made their average mean estimation significantly better than was noted in Herdener et al. (2016). While other studies published by our research team have used the SPUN task to examine mean estimations, they incorporated significant deviations from the methodology used here, including using curved trajectories (experiment 2 in Herdener et al., 2016), adding background visualizations such as "variability clouds" (Pugh et al., 2018, Herdener et al., 2019), and emphasizing specific parts of the instructions (Herdener et al., 2018). As such, additional comparisons are not valid here.

Overall, while average performance when estimating spatial means still was not at an optimal level based on the underlying mathematics inherent to the SPUN paradigm, participants in the current study did reach a better level of performance compared to the most closely related SPUN study. Overall, the spatial mean results presented here support the conclusion that people can accurately learn and respond to average spatial trends, but error does increase with angle and speed variability.

CHAPTER 8 - METHODS: DIGIVAR

To assess understanding of numeric means and variability, we developed DigiVar to incorporate procedures analogous to those of SPUN but by using semi-randomized lists of numbers as stimuli. Experimenters administered this task through Qualtrics (2019). The task was structured with one practice block and four experimental blocks. In each experimental block, participants viewed twenty semi-random two-digit numbers and estimated an ongoing mean after each number. At the end of each number list, participants estimated the variability of all list numbers (See *Figure 17*).

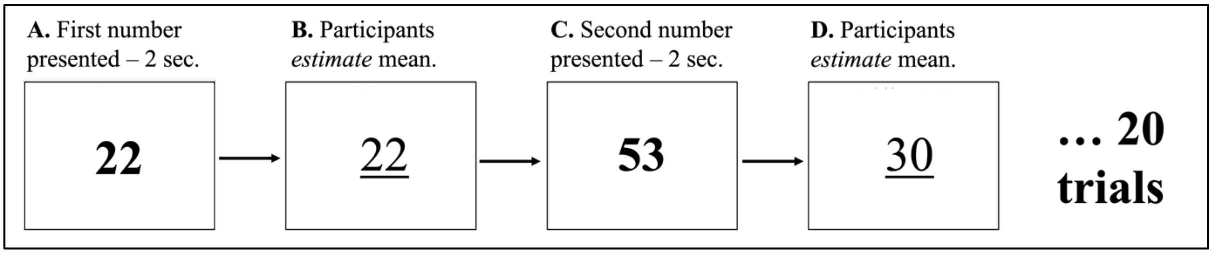
A. First number presented – 2 sec.
B. Participants *estimate* mean.
C. Second number presented – 2 sec.
D. Participants *estimate* mean.
22
22
53
30
… 20 trials

Figure 17. DigiVar Example. The DigiVar paradigm presents 20 semi-random two-digit numbers, and the participant estimates an ongoing mean after each number.

Mean Estimation

During every experimental block, participants saw twenty numbers, each presented for 2 seconds. After each number was presented, participants estimated the mean of all numbers they had seen in that set (i.e., that number and all previous numbers in that list). Participants only viewed each number once. As such, they were required to build an accurate, but estimated, mean over all 20 numbers in each list. Each mean estimation in DigiVar was analogous to each mean trajectory estimation in SPUN. The total time for the 20 trials of Phase 1 was approximately 40 seconds, given the 2 second presentation, and approximately 2 seconds taken to type in the 2-digit response.

Variability Estimation

After viewing all 20 numbers and estimating the final average, participants were immediately asked to consider all numbers they had seen in the current list. Using two sliding number scales, each bounded by 1 and 100, participants were instructed to provide a range of numbers (minimum and maximum) that encompassed approximately 70% of the numbers they had seen in that list (See *Figure 18*). This exercise probed participants' implicit understanding of the numeric variability in each list and is directly analogous to the adjusted circle employed in Phase 1 of SPUN, which probed spatial variability.

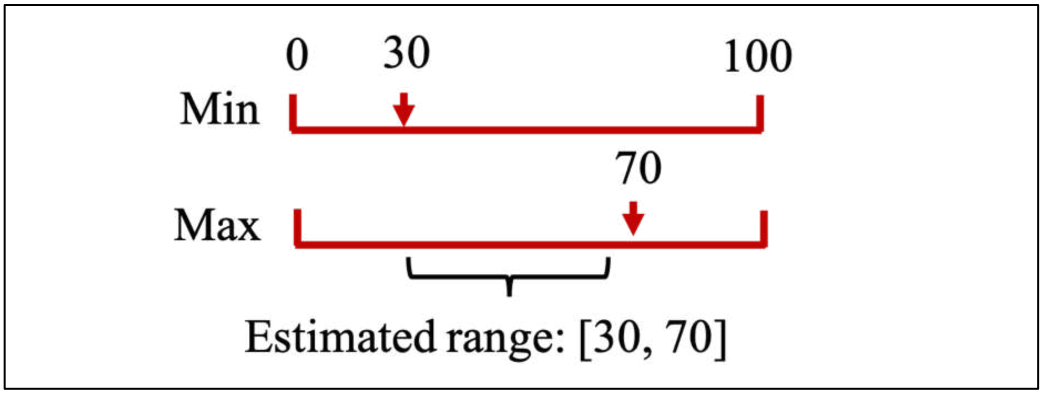

Figure 18. DigiVar Range Estimation. After estimating 20 means, participants provide a minimum and a maximum that captures approximately 70% of the numbers in the list.

Key Variables

Participants estimated variability by providing a number range that captured 70% of the numbers they saw. We calculated how many numbers from each list fell into participant's estimated ranges and treated this as a singular variability estimate. Regarding *sensitivity to numeric variability*, we subtracted the average numbers in range for the low-mean low-variability condition from the high-mean high-variability condition. *Numeric variability bias* was obtained by subtracting the true digits in range from the estimated digits in range. Because each list contained 20 numbers, the true digits in range was a constant 14 digits. Perfect calibration to variability occurred if participants reported intervals that contained 14 list numbers. Positive bias values suggest overestimation of numeric variability, whereas negative bias values suggest underestimation of numeric variability. Like above, *absolute numeric variability error* was the absolute value of bias.

Regarding estimated averages, we again derived a measure of *absolute mean error*; this is the absolute value of the difference between participants' estimated averages and the true averages in their final mean estimate for each list.

Number Lists

Four lists of twenty numbers were each generated with differences in mean and variability (i.e., standard deviation). All digits ranged from 1 to 100. Participants saw the four lists in a fixed order, but numbers were presented randomly. We manipulated the mean and standard deviation of each list such that participants experienced all combinations of high-low mean and high-low variability. The conditions are henceforth labeled:

- HM HV (high mean [M = 61.35], high variability [σ = 22.19]),
- HM LV (high mean [M = 71.40], low variability [σ = 9.85]),
- LM HV (low mean [M = 41.10], high variability [σ = 17.63]), and
- LM LV (low mean [M = 52.20], high variability [σ = 14.49]).

CHAPTER 9 - RESULTS: DIGIVAR

Variability Estimation

Distribution of responses: Numeric variability. Participants responded to numeric variability by providing a minimum and a maximum value that encompassed what they judged to be 70% of the numbers they saw. From this minimum and maximum, we calculated a range for each condition, and then calculated how many numbers in each list fell within participant's given ranges (i.e., digits in given range). *Figure 19* presents the distribution of the digits in given range variable. Responses for each participant were averaged across conditions. The histogram reveals that most frequently, average estimates ranged between 5 and 20 digits in range.

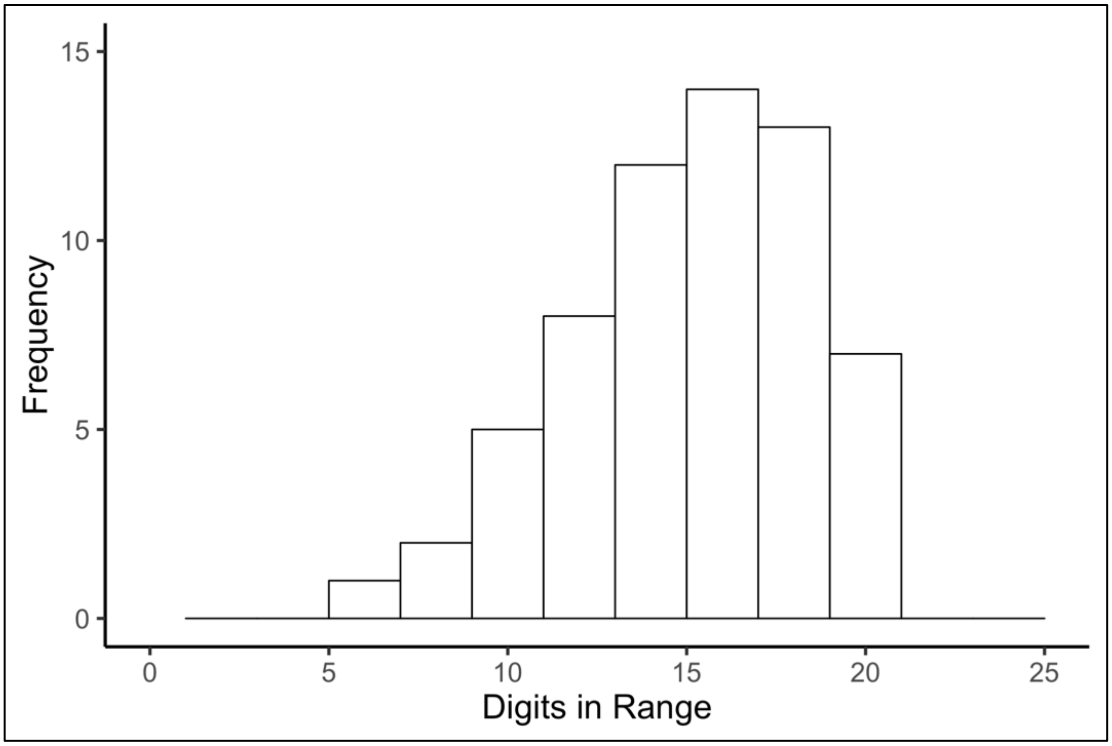

Figure 19. Histogram of Digits in Range. Responses for each participant were averaged across conditions. Average estimates ranged between 5 and 20 digits in range.

Numeric variability sensitivity. Participants responded to numeric variability by providing a minimum and a maximum value that encompassed what they judged to be 70% of the numbers they saw. From this minimum and maximum, we calculated a range for each condition, and then calculated how many numbers in each list fell within participant's given ranges (i.e., digits in given range). In each of the four manipulated conditions, the true, sensitive, 70% range would have contained 14 numbers, since there were 20 presented. *Figure 20* presents the average estimated and true digits in given range each condition.

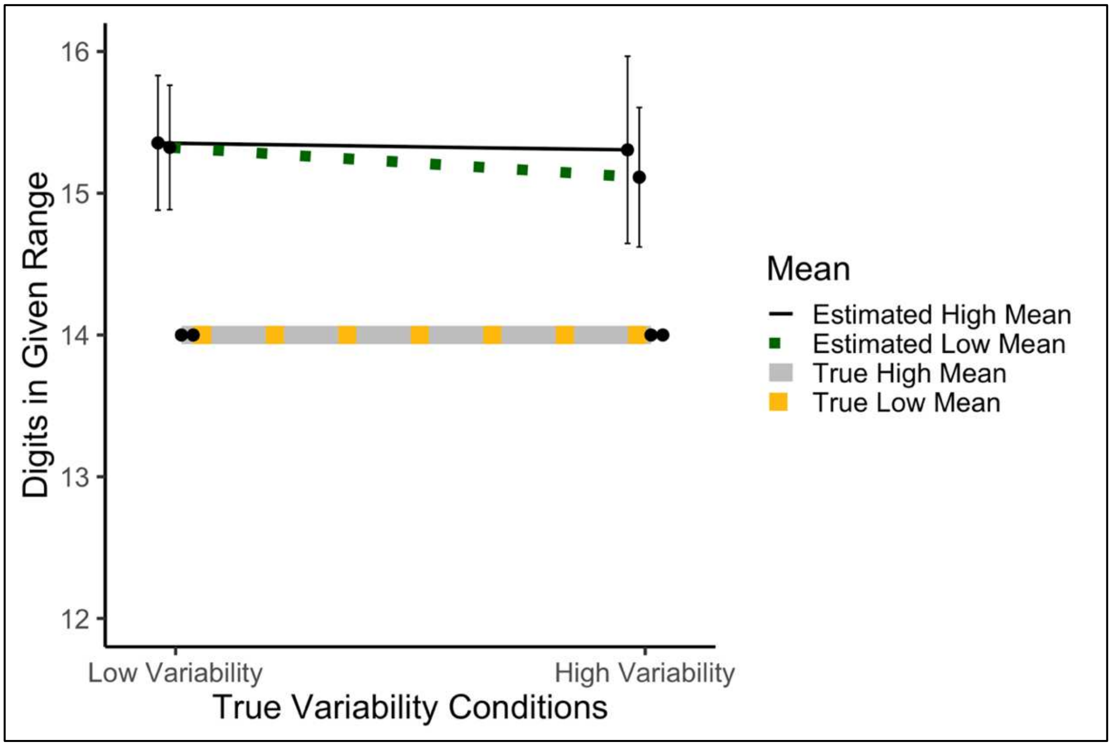

Figure 20. Sensitivity to Numeric Variability [Digits in Range]. On average, the digits in given range variable were similar across all conditions.

A 2X2 MANOVA indicated that neither the main effect of true list mean [Wilk's Λ = 0.99, F(1, 60) = 0.1, p = .75, η_p^2 = 0.002], list variability [Wilk's Λ = 0.99, F(1, 60) = 0.09, p = .76, η_p^2 = 0.002], nor their interaction had a significant effect [Wilk's Λ = 0.99, F(1, 60) = 0.08, p = .77, η_p^2 = 0.01] on average digits in given range were significant. No order effects were noted for this variable.

Numeric variability bias. To determine participant bias to numeric variability, we subtracted the true constant of 14 digits from each participant's digits in given range variable.

This operation produces a signed bias variable that identifies the extent to which participants overestimated variability (i.e., participant's ranges encompassed more than 14 numbers) or whether they underestimated variability (i.e., participant's ranges encompassed less than 14 numbers). *Figure 21* presents the average numeric bias for participants for all conditions.

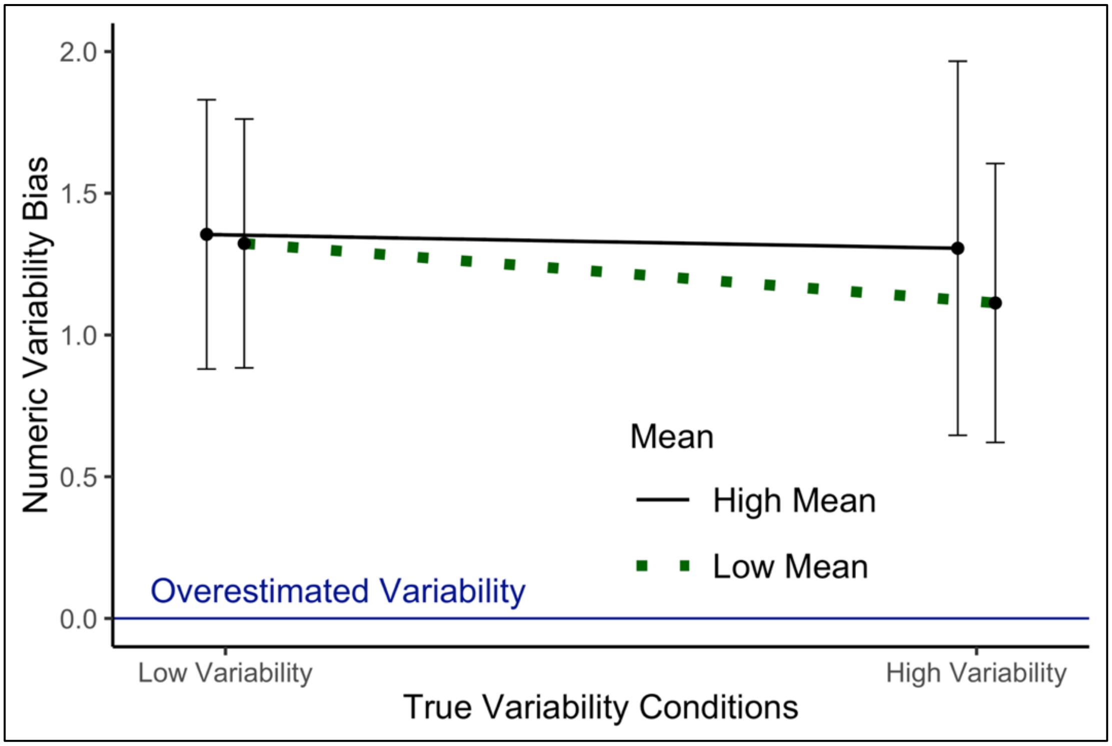

Figure 21. Numeric Variability Bias. On average, participants overestimated numeric variability for all conditions.

Figure 21 shows that for all conditions, participants, on average, overestimated numeric variability. A 2X2 MANOVA indicated that neither the main effect of true list mean [Wilk's Λ = 0.99, F(1, 60) = 0.1, p = .75, η_p^2 = 0.002], list variability [Wilk's Λ = 0.99, F(1, 60) = 0.09, p = .76, η_p^2 = 0.002], nor the interaction [Wilk's Λ = 0.99, F(1, 60) = 0.08, p = .77, η_p^2 = 0.01] was significant.

Given prior literature which detailed how people typically underestimate numeric variability, Hypothesis 3 posited that they would do so on this task as well. Clearly, participants overestimated variability, but to test this, we ran a one-sample t-test on average numeric bias using a test value of zero (no bias). Results confirmed that participants, on average, significantly overestimated variability [$t(61) = 3.05$, $p < .01$, $d = 0.38$].

Variability estimates by block. We also examined whether any differences occurred between participants' variability bias between trials. *Figure 22* shows the average bias for each trial.

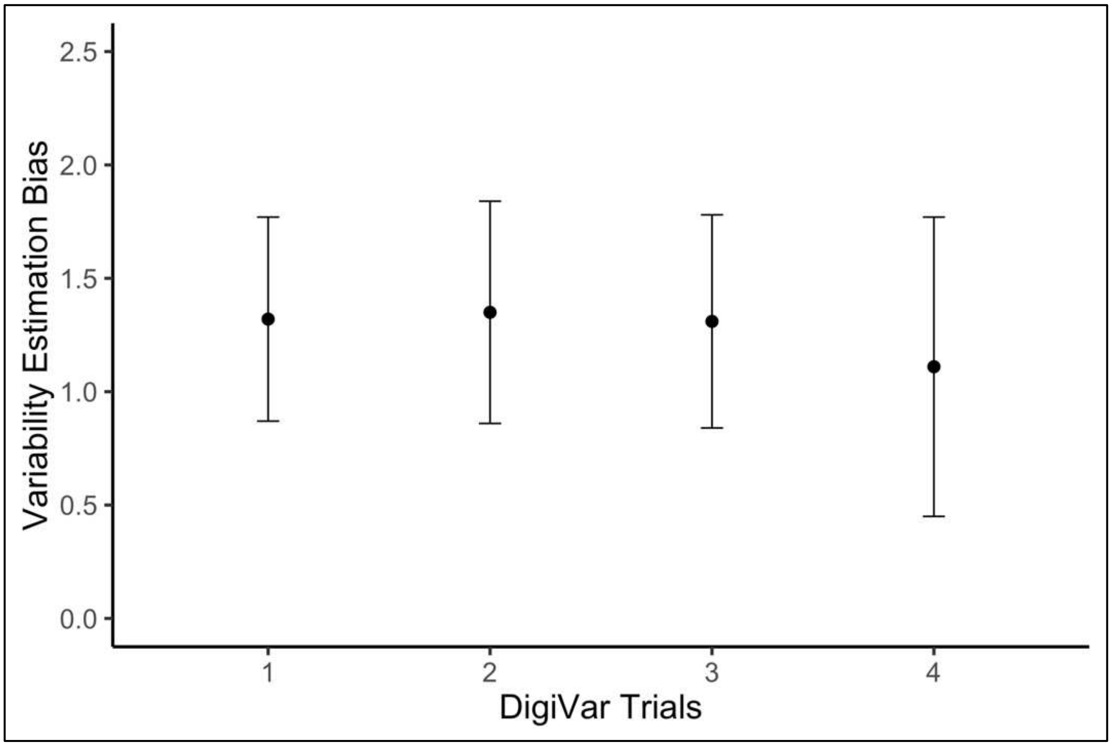

Figure 22. Numeric Variability Bias by Trials. This figure shows how average bias to numeric variability changed across trials. No significant differences occurred between trials.

This figure demonstrates the similar bias expressed in each block. No significant differences occurred between average bias in each trial ($p > .05$).

Numeric variability bias and task order. Regarding the effect of task order on bias, no effects were significant except the three-way interaction between list mean, list variability, and task order [Wilk's $\Lambda = 0.92$, $F(1, 60) = 5.31$, $p = .03$, $\eta_p^2 = 0.08$]. *Figure 23* presents the average numeric variability bias for those participants who completed the SPUN task first and *Figure 24* presents the average numeric variability bias those who completed the DigiVar task first.

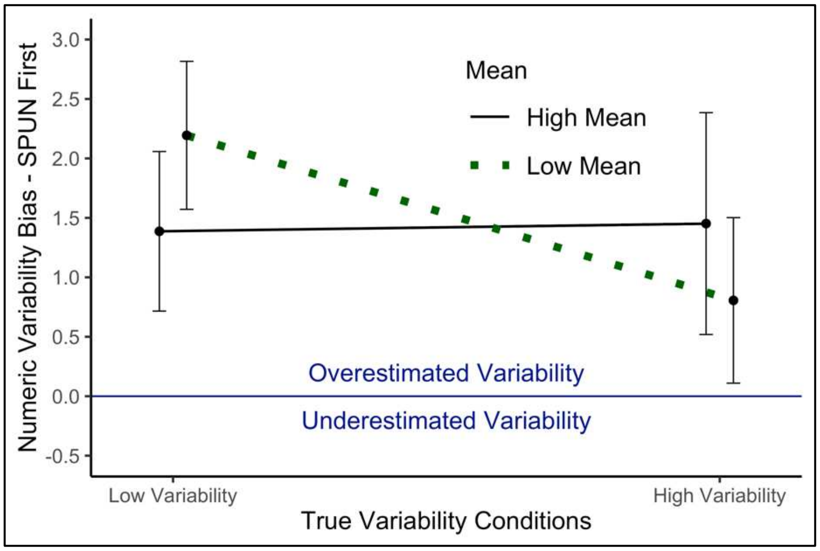

Figure 23. Numeric Variability Bias for those participants who completed the SPUN Task first. When the SPUN task was first, higher true variability with a lower mean led to less numeric variability overestimation.

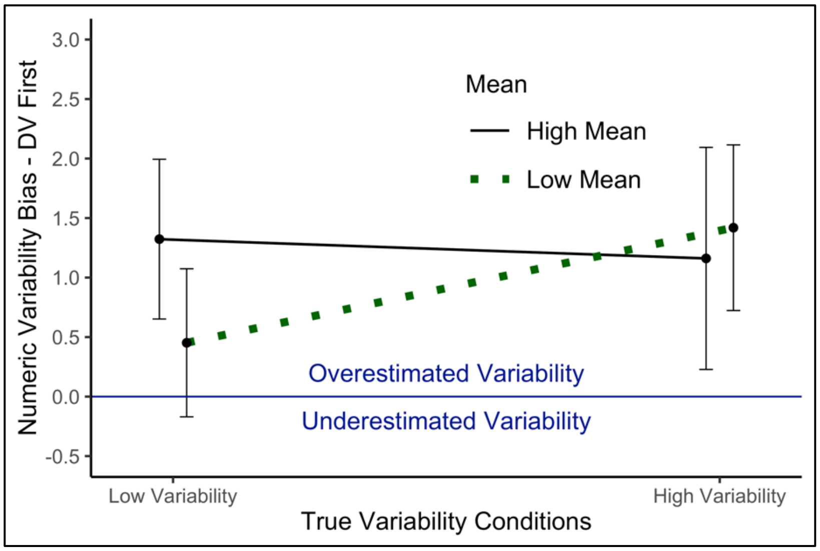

Figure 24. Numeric Variability Bias for those participants who completed the DigiVar Task first. When the DigiVar task was first, higher true variability with a lower mean led to more numeric variability overestimation.

This interaction reveals that when the SPUN task was first, higher variability with a lower mean led to less numeric overestimation, but when the DigiVar task was first, this estimated variability pattern reverses, which cannot be easily explained.

Absolute numeric variability error. *Figure 25* presents the numeric absolute variability error across conditions. Larger values suggest *more* error.

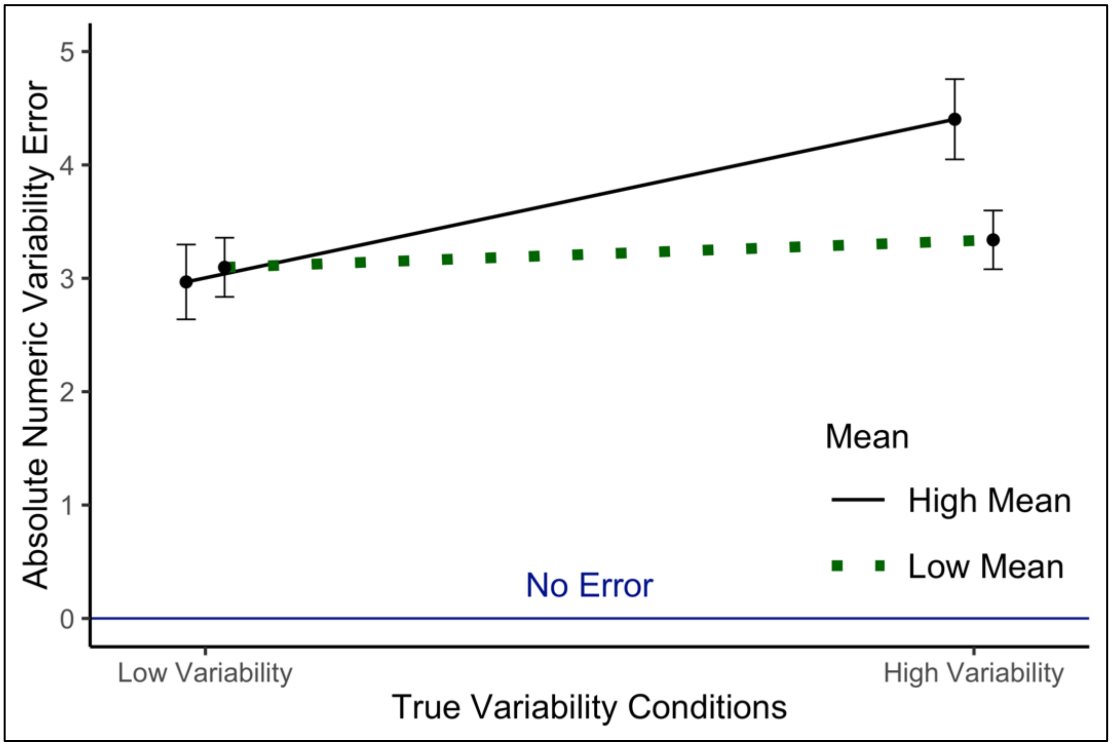

Figure 25. Absolute Numeric Variability Error. On average, as true list variability increased, absolute numeric variability also increased.

As seen above, as true list variability increased, absolute numeric variability also increased. This effect was more pronounced for the high mean condition. A 2X2 MANOVA revealed a significant main effect of list variability on estimated variability precision [Wilk's Λ = 0.90, F(1, 60) = 6.98, p = .01, $\eta_p{}^2$ = 0.10], an effect that is most meaningfully interpreted in the context of the significant interaction between list mean and variability [Wilk's Λ = 0.91, F(1, 60) = 5.71, p = .02, $\eta_p{}^2$ = 0.09]. Specifically, when the list had a high mean, an increase in variability led to an increased absolute error when estimating list variability. This variable was not affected by task order.

Mean Estimation

Distribution of responses: Numeric mean error. Similar to SPUN, we calculated participants' absolute numeric mean estimation error. *Figure 26* presents the frequency of mean error averaged across all conditions. The histogram reveals that most frequently, numeric mean error ranged from 0 to 20.

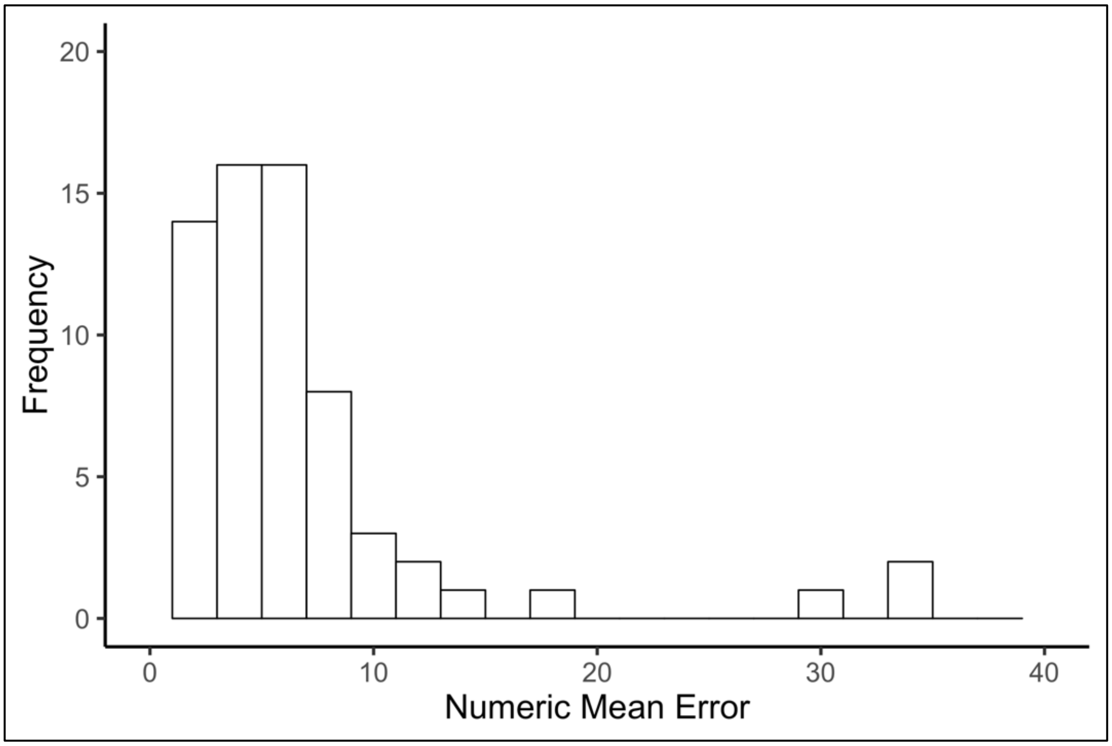

Figure 26. Numeric Mean Error. This figure presents participants' numeric mean error, averaged across participants. Most frequently, numeric mean error ranged from 0 to 20.

Absolute numeric mean error. Similar to SPUN, we calculated participants' absolute numeric mean estimation error. Across all conditions, average absolute error was 6.45 (SD = 5.61), meaning that, on average, participants' mean estimates were approximately 6.5 digits deviant from the true mean. *Figure 27* presents this variable for each condition.

Data in *Figure 27* demonstrate that, as true list variability increased, absolute mean estimation error increased. This trend was more pronounced for the high mean conditions compared to the low mean conditions.

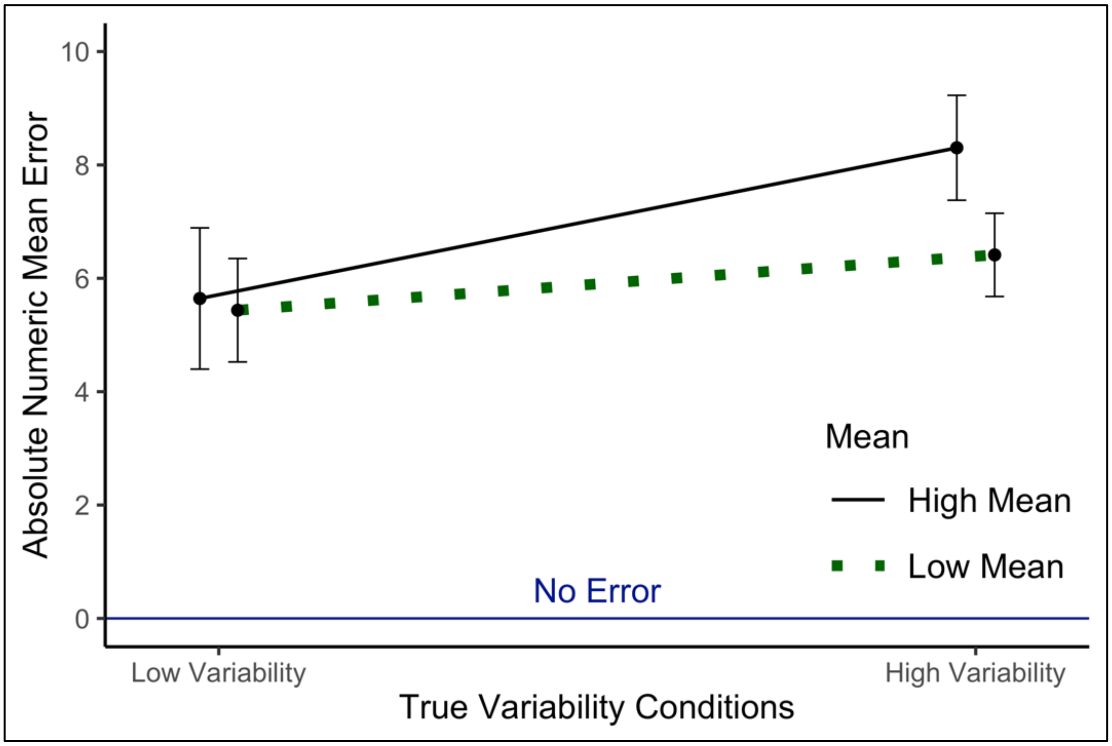

Figure 27. Absolute Numeric Mean Error. As true list variability increased, absolute mean estimation error increased.

A 2X2 MANOVA revealed that the main effect of list mean was not statistically significant [Wilk's Λ = 0.96, F(1, 60) = 2.52, *p* = .12, η_p^2 = 0.04]. The main effect of list variability was trending towards significance [Wilk's Λ = 0.95, F(1, 60) = 3.16, *p* = .08, η_p^2 = 0.05]. Moreover, the interaction between list mean and list variability was also trending towards significance [Wilk's Λ = 0.95, F(1, 60) = 3.03, *p* = .09, η_p^2 = 0.05], as seen by the divergence of the low mean (i.e., dotted green) and high mean (i.e., solid black) lines in *Figure 27*. No significant order effects were noted for this variable.

CHAPTER 10 - DISCUSSION: DIGIVAR

Task Order

As in the SPUN analyses, a task order variable was included in the MANOVAs. The only significant effect involving task order was a three-way interaction with true list mean and variability on participants' numeric variability bias. The lack of significant task order main effects suggests that completing the DigiVar task first made little impact on performance. This is the extent to which we will discuss the influence of task order on DigiVar results.

Sensitivity to Numeric Variability

The conceptual diagram *Figure 28* below presents trends in estimated and true variability in the DigiVar task. The x-axis represents true variability, and the y-axis represents estimated variability. The black line represents perfect calibration, and the dotted green line represents the average responses to variability. We did not calculate a sensitivity ratio for DigiVar because perfect calibration corresponded to 14 digits both at low list variability and high list variability (leading to a denominator of zero in a ratio). This figure demonstrates that participants were sensitive to changes in variability, seen through the similar positive slopes of the green dotted line and the solid black line. These findings of proper sensitivity to true variability are contrary to the variability sensitivity presented in Beach and Scopp (1968); in their experiment, at low levels of true variability, participants were actually slightly hypersensitive. At higher levels of true variability, participants were hyposensitive. This study, conversely, found that participants showed fairly calibrated sensitivity; at both low and high list variability participants gave similar, if overestimated estimates of variability (see *Figure 21*). These findings suggest that participants were somewhat aware of the changing variability between numbers in the lists, and they adjusted their 70% intervals accordingly.

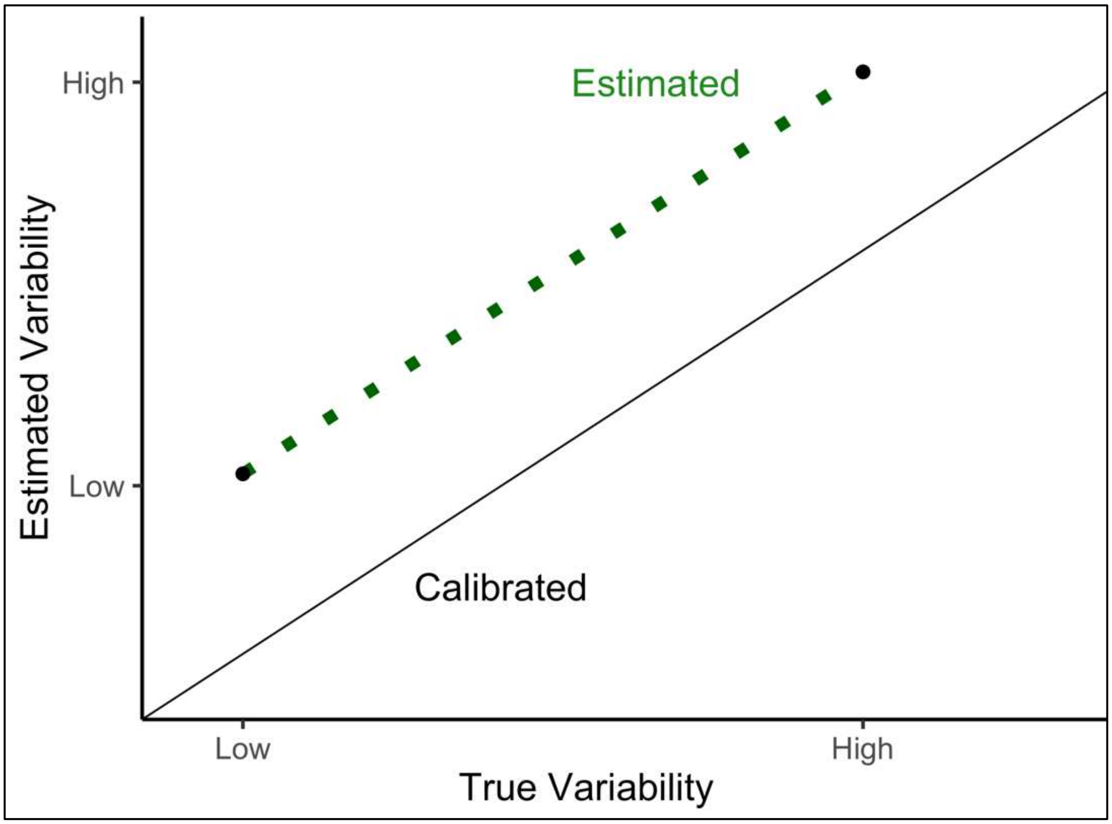

Figure 28. DV Sensitivity Diagram. This diagram shows how estimated variability changed as a function of true variability in DigiVar. Participants showed fairly calibrated sensitivity; at both low and high list variability participants gave similar, if overestimated estimates of variability.

Variability Bias

For each condition in DigiVar, calibrated variability would be providing a range of 14 numbers, which means that participants slightly overestimated variability (seen by the green dotted line being slightly higher than the black calibration line). Participants overestimated variability at a similar level for all conditions, with an average of 9.1%. Such a low amount of overestimation suggests that participants were fairly well calibrated to numeric variability. This trend could be due, in part, to the familiarity of numeric stimuli, a tenant of the MOVE model. The level of stimuli familiarity primarily impact encoding of variability, with more familiar material leading to more accurate mental models of that stimuli. The high familiarity of numbers is without question; indeed, evolutionary research suggests that "numerical competencies" are preverbal, with infants responding to differently to varying number arrays in habituation paradigms (Tosto et al., 2014). Arguably, then, the high familiarity of numeric stimuli ease encoding of numeric variability, leading to mostly calibrated estimates found here.

The overestimation of variability noted here contrasts with a key finding from Beach and Scopp (1968), one of the few studies to examine estimates of numeric variability. These researchers found that participants underestimated the variability of normal distributions; they subsequently hypothesized that because most deviations in such a distribution are close to the mean, participants emphasize these small deviations and thus underestimate variability. Furthermore, using various stimuli, multiple studies noted that estimates of variability decrease as the mean of the sample increases (Hofstatter, 1939; Lathrop, 1967; Beach and Scopp 1968). These experimenters, however, had participants make single judgements (e.g., "more variable" versus "less variable") as opposed to providing a point estimate or range for variability, as was done in the present study. This difference in the response methodologies is also an integral part of the MOVE model, and certainly may account for the bias dissimilarity noted between this study and others that note underestimation of variability.

Another plausible explanation for the clear overestimation of variability comes from the instructions for participants to respond to variability. After estimating all means for each list, participants were given the following directions: "Consider all of the numbers that you saw from this group. Please provide a range numbers (minimum and maximum) that encompasses approximately 70% of the numbers you saw." In retrospect, participants could have approached the task with the goal of *at least* capturing 70% of the numbers they saw but giving slightly larger estimates to be positive that they provided a large-enough range. Additional research needs to follow up on this aspect of the DigiVar task, exploring different response techniques and their impact on variability judgements.

Estimating Numeric Means

For each 20-digit number list, participants estimated a "running" mean after each number. For every list, we examined the bias and precision of the final estimated mean to identify if these estimates were influenced by the magnitude of the true list mean and variability. Results suggests that participants' mean estimates were sensitive to changes in both the true list mean and variability.

The precision results presented above are best compared to existing literature focused on the estimation of numeric means. Beach and Swenson (1966)'s subjects provided highly accurate estimates of numeric means, with an average absolute error of 3.51 digits. Further, as the number of list numbers in the lists increased, mean estimates became less precise. Across all conditions, participants in the current study had an average absolute mean estimation error of 6.45 (SD = 5.61), and while this average is greater than Beach and Swenson, they had a maximum number of seven digits in their lists, where we had 20 in each list. As Beach and Swenson suggest that more numbers in the list lead to increased error, it stands to reason that we found a higher average estimation error.

Beach and Swenson (1966) also found that increasing the variability within the number lists corresponded to less precise mean estimates. While the main effect of list variability was only trending towards significance (see *Figure 27*), average absolute mean estimation error did increase as variability increased. Finally, Taken together, these results suggest that when estimating a numeric mean, as the variability between numbers or numbers in the list increased, participants had a more difficult time accurately updating their mental model of the average. Moreover, knowing that they would be required to provide an estimation of variability at the end of each list may have encouraged them to attend to the variability, at the expense of the mean estimates.

CHAPTER 11 - METHODS: CORRELATIONAL ANALYSES

The main aim of this project was to identify the extent to which participants expressed similar variability estimation performance across numeric and spatial stimuli. Initially, we had thought to correlate bias and precision, but have since determined that sensitivity, rather than precision, better captures the ability to estimate different levels of variability. This is because precision, as a variable, is confounded by sensitivity: if a participant shows no sensitivity to variability, a precision value can still be calculated, but its value will depend entirely on the level of variability assessed and on the underlying bias (See *Figure 2*). As such, we revised this section to include correlations using sensitivity to variability, as opposed to precision.

The three key variables used in these correlations include variability sensitivity, variability bias, and error in estimating or predicting the mean. For the correlations, sensitivity is the same as above; that is, the variable was calculated by subtracting the estimated variability at low true variability from that at high true variability. To correlate bias, we calculated an average bias expressed by participants across the conditions in each task (low and high variability for speed and angle in SPUN, and low and high mean and variability in DigiVar). A similar average variable was calculated for mean estimation error.

The parametric assumptions of linearity and normality were checked prior to running these analyses. We utilized the Pearson's Product-Moment Correlation when variables met all the parametric assumptions, and when they did not, we utilized Spearman's Rho. To evaluate the strength of the correlations, we used the effect size cutoffs presented in Cohen (1992).

CHAPTER 12 - RESULTS: CORRELATIONAL ANALYSES

Variability Sensitivity

No significant between-task correlations occurred for variability sensitivity variables. This was true for both the correlations of SPUN Phase 1 and DigiVar sensitivity (r_s = -0.03, p = .81) and SPUN Phase 2 and DigiVar sensitivity (r_s = -0.08, p = .52). These correlations suggest that the ability to perceive variability in the spatial domain is unrelated to perceiving variability in the numeric domain.

The correlation between variability sensitivity measures in SPUN Phase 1 and SPUN Phase 2 was statistically significant (r_s = -0.28, p = .03); while this is a negative correlation, because of the way Phase 1 and Phase 2 were measured, it means that those who were more sensitive in Phase 1 were also more sensitive in Phase 2. As such, there is some evidence that Phase 1 and Phase 2 assess a commonality in the ability to perceive and respond to spatial variability. This will be discussed further below.

Variability Bias

A moderate, positive, significant correlation (r_s = +0.30, p = .02) occurred between average SPUN Phase 1 and DigiVar variability biases, suggesting that those who demonstrated a bias when adjusting a circle also demonstrated a similar bias when adjusting a numeric interval.

The correlation between SPUN Phase 2 and DigiVar variability biases was not statistically significant (r_s = +0.22, p = .09) and the correlation between SPUN Phase 1 and SPUN Phase 2 variability biases was not significant (r_s = -0.03, p = .80).

Estimation Error of the Mean

A large, significant, positive correlation (r_s = +0.44, $p < .01$) occurred between estimation error of the mean in the SPUN task and the DigiVar task. This suggests that those who estimated spatial means with more error also estimated numeric means with more error.

Estimation Error of the Mean and Variability Sensitivity

For the SPUN task, the correlation between participant's average estimation error of the mean and Phase 1 variability sensitivity was moderate and significant (r_s = -0.38, $p < .01$). This means that those participants who estimated spatial means with less error were more sensitive to spatial variability. The correlation of average mean estimation error and Phase 2 variability sensitivity was not significant (r_s = +0.22, $p = .09$).

For the DigiVar task, the correlation of participant's average estimation error of the mean and variability sensitivity was not significant (r_s = +0.10, $p = .43$).

CHAPTER 13 - DISCUSSION: CORRELATIONAL ANALYSES

Correlations between key variables formed the core of this series of experiments. The following discusses implications of these relationships.

Variability Sensitivity

For sensitivity correlations, the only significant relationship occurred between sensitivity measures in SPUN Phase 1 and Phase 2. Of note, how we manipulated variability across the two Phases was comparatively opposite. For example, sensitive responses to large true variability in Phase 1 would result in a *larger* circle but would result in a *smaller* estimated variability in Phase 2. If trajectory endpoints are highly variable, then a smaller percentage fall in the fixed circle probe. The derived variability sensitivity measures for each participant in Phase 1 and Phase 2 retain this opposite valence. Thus, we can interpret the significant negative correlation as so: those individuals who demonstrated more sensitivity to variability in Phase 1 also demonstrated more sensitivity in Phase 2. This suggests that, to a small extent, Phase 1 and Phase 2 assessed a similar sensitivity to spatial variability unrelated to the response method (i.e., adjustment versus assessment). This finding supports the perspective of individual differences in sensitivity to spatial variability – some participants demonstrated more sensitivity in both SPUN Phases but did not retain this performance for numeric stimuli as well. The almost absent correlations between variability sensitivity measures in SPUN and DigiVar means that participants did not demonstrate domain-general sensitivity to variability.

Variability Bias

A moderate, positive, significant correlation occurred between biases in SPUN Phase 1 and DigiVar, meaning that those participants who gave larger circle sizes in SPUN also gave larger numeric intervals in DigiVar. This correlation calls out the analogous response methodology used for both probes was analogous: that of adjusting a subjective confidence interval.

Correlations involving the direct estimate of spatial variability in SPUN Phase 2 were not statistically significant. This could have been due to the difference in response methodologies – the SPUN Phase 1 and DigiVar adjustment does not correspond to the direct assessment of variability used in SPUN Phase 2. These findings provide evidence for the effect of response methodology on variability bias as described by MOVE (Wickens et al., 2020). Herdener et al., (2019) noted a similar dissociation between adjustment and assessment in SPUN.

Estimation Error of the Mean

A large, significant, positive correlation occurred between mean estimation error in DigiVar and SPUN, suggesting that those individuals who estimated spatial means with more error also did so with numeric means. This implies a general ability of individuals to estimate average behavior across stimuli.

Individual differences in the ability to estimate such central tendencies could be due to various factors. Those people who are exposed to more average behavior (e.g., average spatial weather patterns, baseball batting averages, average earnings per year) and who have more practice making average estimates may have developed a general, cross-stimuli ability to make such estimates. Additionally, insofar as estimating averages is a form of categorization (e.g., classifying new instances based on averaging features; Smith, Zakrzewski, Johnson, Valleau, & Church, 2016), those people better at categorization may retain an advantage when estimating any average behavior. Indeed, individual differences have been noted in studies of categorization, with a link to working memory capacity as a mediating factor (Lewandowsky, 2011). Additional research is needed to identify why people differ in the error of their average estimates.

Estimation Error of the Mean and Variability Sensitivity

Regarding the relationship between estimation error of the mean and variability sensitivity, those individuals that estimated means with less error demonstrated more sensitivity to variability in SPUN Phase 1. This was not anticipated – using the SPUN paradigm, Herdener et al. (2019) found an attentional tradeoff, where individuals who estimated means with more error were more accurate when estimating variability, as if they could attend to either estimating the mean, or estimating variability, but not both.

This finding can be explained in at least two respects. First, this could suggest a common spatial ability, such that individuals show similar performance for spatial probes throughout the task. While the correlation between estimation error of the mean in Phase 1 and spatial variability assessment in Phase 2 was only marginally significant, coupled with the significant Phase 1 mean and variability sensitivity correlation, evidence exists of general individual differences in spatial ability.

Participant's level of attention to the task could have also influenced these results. If some participants attended more to the SPUN task, then they would be more likely to perform well compared to others who didn't focus as much. Indeed, this is certainly a risk for MTurk samples, as researchers cannot control the experimental environment. Additional laboratory research is needed to see if attention plays a significant role in these findings.

CHAPTER 14 – GENERAL DISCUSSION

The current study utilized an individual differences approach to investigate the extent to which people have a domain general ability to understand the variability and average behavior of continuously distributed stimuli. Using SPUN and a newly developed analogous numeric task, we assessed sensitivity to different levels of variability and bias/precision of responses to variability/means. Here, we briefly discuss the magnitude of support for each hypothesis.

Hypothesis 1A. A significant between-task correlation will occur between the sensitivity of variability estimations. No significant correlations occurred between variability sensitivity in SPUN and DigiVar – those who were more sensitive to different levels of spatial variability did not retain this sensitivity for numeric variability. This suggest that multiple cognitive processes/systems likely contribute to the understanding and estimation of variability for different types of stimuli.

We can interpret the lack of between-task correlations in terms of storage in different memory systems. For instance, while some researchers have found a significant correlation between individual's spatial and verbal working memory (Unsworth, Brewer, & Spillers, 2009), this still leaves some differences unaccounted for in storage capacity. Perhaps sensitivity to variability in different types of stimuli are served by different subsystems of working memory: numeric variability by verbal working memory and spatial variability by spatial working memory.

Another possible factor for the failure to find between-task correlations lies in the amount of variability manipulated in DigiVar versus SPUN. More total variability was manipulated in SPUN, meaning that the sensitivity measure covered a larger amount of variability as well.

Additional research that equivalates the amount of variability manipulated within-tasks will elucidate the extent to which this explains the current findings. This will be discussed further in the limitations and future directions section below. Overall, this hypothesis was not supported.

Hypothesis 1B: A significant between-task correlation will occur between variability bias (signed error) performance. The correlation of average variability bias in SPUN and DigiVar tasks was positive and significant only when using adjustment as a response method. This is an important distinction, because adjustment taps into more of the perceptual response to variability, as in the System 1 quick, heuristic responding. Perception occurs more automatically and can be seen as a part of System 1 cognition. The slower, System 2 cognitive process is responsible for responses like variability assessment, because such responses necessitate careful conversion from viewed variability to a numeric form (Kahneman, 2011). Evidence from this study suggests that variability bias may stem primarily from the perceptual-cognitive adjustment of spatial or numeric intervals. This hypothesis was partially supported.

Hypothesis 1C: A significant between-task correlation will occur between estimation error of the mean. Correlation results suggest that those who estimated spatial means with more error also estimated numeric means with more error. Participants varied in their amount of estimation error. Possible drivers of the individual differences noted here include working memory capacity and the ability to categorize. Exposure to, and practice estimating, average trends in continuous stimuli behavior likely also plays a role in making such judgments. Additional research is needed to clarify why people demonstrate similar bias across tasks. This hypothesis was supported.

Hypothesis 2: No significant within-task correlations will occur between the variability sensitivity and estimation error of the mean. The correlation between estimation error of the mean in SPUN Phase 1 and variability sensitivity in SPUN Phase 1 (adjustment) was significant, suggesting that those participants who estimated spatial means with less error were more sensitive to spatial variability. This was unanticipated based on our research group's prior studies, which noted a possible attentional tradeoff between mean and variability probes in Phase 1 (Herdener et al., 2018, 2019).

For DigiVar, the correlation between absolute variability and estimation error of the mean was not significant, suggesting that people who were more accurate on mean estimates were not more accurate with their variability estimates. This was likely not due to an attentional tradeoff, as mean and variability estimates occurred sequentially. Additional research is needed to identify why the numeric task did not result in the similar trends seen in the spatial task. This hypothesis was partially supported.

Hypothesis 3: DigiVar task performance will reflect underestimation of variability. In the in the DigiVar task, participants significantly overestimated variability via their number ranges. This was inconsistent with prior literature (Beach & Scopp, 1968; Henrion & Fischoff, 1985; Hansson, Juslin, & Winman, 2008), but may have been due, in part, to interpretation of the instructions. This will be discussed further in the limitations and future directions section.

Individual Differences

Estimation bias. While they focused on overconfidence, Pallier and colleagues (2002) did find individual differences in the accuracy of confidence judgements. Trends in the current study also suggests a role of individual differences in bias to variability, but the bias elicited was overestimation of variability. We didn't test any possible mediators of the ability to estimate variability, but this is an important next step. As described in MOVE (Wickens et al., 2020), possible abilities that may drive a similar bias to spatial and numeric variability include personality traits towards being over- or under-confident (Pallier et al., 2002). If underestimation of variability is overconfidence in the estimation of the mean (Henrion & Fischoff, 1985; Wickens et al., 2020), then the converse implies that overestimation of variability is underconfidence in the estimation of the mean. This possibility needs additional exploration. Other cognitive mediators that could influence these results include working memory (Hansson et al., 2008), and numeracy (Rinne & Mazzaco, 2013).

This evidence of individual differences in variability bias has important implications for human factors applications. The same bias was elicited across stimuli, and thus similar interventions may be used to help reduce bias. Identifying when biases are most likely to occur will allow for proper human factors interventions, such as visualizations and training on domains that include uncertainty, such as extreme weather patterns.

This study affords some interesting caveats to the MOVE model of variability estimation (Wickens et al., 2020). For one, the SPUN task, which influenced much of the model consistently demonstrated that people overwhelmingly underestimated variability. Findings presented here demonstrate that this bias can apparently be pushed to general overestimation for both spatial *AND* numeric stimuli. The differences that could have caused this change in SPUN were the simultaneous mean/variability judgements in Phase 1, and the bonus incentive to attend to variability. This was the first time that the DigiVar task was tested, so additional experiments are needed to identify if this overestimation of variability can be manipulated.

Estimation error of the mean. Individual differences were also noted for estimation error of the mean across SPUN and DigiVar. While these trends were noted, people did not all perform in a similar manner; some people showed less estimation error of the mean, and others showed more estimation error of the mean across tasks. These individual differences suggest the potential role of mediating cognitive abilities that support estimation of average behaviors but could also imply varied levels of attention to the task – more versus less engagement throughout the tasks can lead to similar results.

Individual differences in mean estimation also have implications for human factors applications. In the realm of decision making, people often make "pseudo-average" estimations by aggregating cues and sets of information. For example, when estimating whether a sports team will win a game is dependent on various sources of data, each which must be aggregated to arrive at a "mean" estimate of likelihood. Additional research is needed to identify the extent to which the ability to estimate averages in this study (e.g., averages from continuously presented numeric/spatial stimuli) are similar to the ability to make discrete outcome estimates from aggregate, or averaging, data. Nonetheless, with similar error elicited across stimuli in the current study, interventions to help people those who struggle more with estimating the mean may help general performance when making such judgments.

Observed Power Analysis

Table 2 presents each of the key between and within-task correlations discussed above and the achieved power for each analysis. These post-hoc power analyses were conducted using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009). These findings demonstrate that the correlations in this study ranged in observed power, with only the two largest correlations achieving adequate power above 0.8.

Table 2. Observed Power Analysis

Correlation	Correlation Coefficient (r_s)	Achieved Power
SPUN Phase 1 and DigiVar Sensitivity	-0.03	0.06
SPUN Phase 1 and 2 Bias	-0.03	0.06
SPUN Phase 2 and DigiVar Sensitivity	-0.08	0.09
DigiVar Sensitivity and Mean Estimation	0.10	0.12

SPUN Phase 2 Sensitivity and Mean Estimation	0.22	0.41
SPUN Phase 2 and DigiVar Bias	0.22	0.41
SPUN Phase 1 and 2 Sensitivity	-0.28*	0.61
SPUN Phase 1 and DigiVar Bias	0.30*	0.67
SPUN Phase 1 Sensitivity and Mean Estimation	-0.38**	0.87
SPUN and DigiVar Mean Estimation	0.44**	0.95

*: Significant at $p < .05$; **: Significant at $p < .01$.

While the correlations between SPUN Phase 1 and Phase 2 Sensitivity and between SPUN Phase 1 and DigiVar Bias were both under-powered, they were still both statistically significant at an alpha level of .05. As such, while a larger sample size may have resulted in more adequate power for these correlations, the interpretation would likely not change much from that stated above.

Limitations and Future Directions

Taken together, these findings provided some evidence that similar cognitive constructs contribute to variability bias across stimuli. Even still, some factors limited these experiments.

Levels of manipulated variability. While participants demonstrated limited sensitivity to increasing variability, only two levels were manipulated in both SPUN and DigiVar. Adding more than two levels of variability in both tasks will be necessary to identify a fuller range of sensitivity to variability across stimuli. Further, the amount of manipulated variability was conceptually different in the two tasks. The SPUN trajectories characterized variability in terms of speed and angle, which created four different distributions of spatial *variability*. Conversely, in DigiVar, numeric variability was manipulated based on the variability of the number lists, which had two levels. This means that the SPUN task covered a wider total range of variability than did DigiVar. Increasing the range of manipulated variability in DigiVar could lead to a different pattern of results – not just stable overestimation of variability (see *Figure 21).*

Financial incentives. Financial incentives also likely influenced the results in this study. As discussed above, this was one of the first SPUN experiments to financially incentivize people (via a bonus) to give calibrated circle sizes in the final four trials of each block. Participants were told that they could earn a bonus during these trials for providing accurate circle sizes, which may have encouraged a different level of processing variability throughout the task. Since participants were only bonused during SPUN and not DigiVar, however, it would be unlikely that this is the driving factor behind a shift from under- to overestimation of variability.

Further experiments should focus on the extent to which financial incentives influence responses to variability, and moreover, if this is a function of attending to the stimuli characteristics underlying that bonus (i.e., participants knew they were going to be bonused for giving accurate responses to variability, so they attend and encode that information better).

Latent model of variability estimation. This study provided some evidence of variability bias across stimuli, specifically when using adjustment as a response technique. Other relationships were not found, suggesting that the ability to understand environmental variability is complex. One future direction for this research is to develop a latent variable model of variability estimation, to see if different patterns emerge from a large-scale set of data. For instance, one could use a latent profile analysis to identify if sets of people aggregate in their abilities to estimate variability (Oberski, 2016). This would facilitate hypotheses like the following: Are there multiple distributions of variability estimation in the population? Does one distribution perform much better on variability estimations than another? Additional grouping variables could be added to such a model that hypothetically explain differences in variability estimation, including numeracy, working memory, and fluid intelligence.

Refining DigiVar. This was the first study to use the newly developed task, DigiVar. This task was developed to be analogous to SPUN, and in some ways, it was (e.g., using the technique of adjustment for variability estimation maps onto the Phase 1 variability probe in SPUN). In future experiments, we will continue testing DigiVar to determine the malleability of responses. For instance, the numbers generated for the lists were semi-random and approximately normally distributed. Using different types of numerical distributions (e.g., highly skewed distributions, saddle-shaped distributions) could make the task more difficult and lead to a different pattern of results. Other ways to change the stimuli in DigiVar could be to significantly increase or decrease the presentation speed, use three- or four-digit numbers, or use negative numbers. These changes will help identify how different characteristics of numbers and distributions influence overall results. Finally, the instructions in DigiVar could have been misinterpreted to provide "at least 70%" of the numbers in each list, leading to the findings of variability overestimation. These instructions should be refined and tested, to ensure that participants accurately understand that the goal is to provide a sensitive range as close to 70% as possible.

CONCLUSION

The ability to estimate variability of continuously distributed stimuli is an often-overlooked and under-researched skill. For example, when purchasing a home, buyers often track the current state of the market, and hope to buy when the prices are low. Financial trends, however, are subject to numeric variability, and buyers who accurately understand the amount of variability in the market will likely walk away satisfied. Overestimating the amount of numeric variability, as seen in this study, may negatively influence such major decisions to buy and sell. If one thinks that prices are likely to be more variable in the future (overestimating variability), they may make different decisions in the present (i.e., if a consumer thinks that a low current price will radically change in the future, they may be more inclined to buy right away before that anticipated change).

The understanding of spatial variability can also be applied to a number of real-world situations. Trajectory of weather patterns are the common application; while the most likely location of a storm may be easy to understand, the amount of variability represented by weather graphics are unintuitive and adversely impact decisions to evacuate during a severe storm. Consequently, proper estimations of variability hold significant implications for both performance and safety.

This study investigated the extent to which participants estimated variability for spatial and numeric information; results suggest that different cognitive mechanisms likely support the understanding and estimation of different types of information (i.e., people who accurately estimate spatial variability were not necessarily more likely to accurately estimate numeric variability), but that a similar bias may occur across stimuli (i.e., overestimation of variability).

Additional research is needed to continue elucidating the mechanisms involved in these mental operations and how we can help people become better calibrated to variability.

REFERENCES

Alley, R. B., Emanuel, K. A., & Zhang, F. (2019). Weather: Advances in weather prediction. *Science, 363*(6425), 342-344. https://doi.org/10.1126/science.aav7274.

Baddeley, A. D. (2001). Is working memory still working? *American Psychologist, 56*(11), 851-864.

Beach, L. R., & Swenson, R. G. (1966). Intuitive estimation of means. *Psychonomic Science, 5*(4), 161-162.

Beach, L. R., & Scopp, T. S. (1968). Intuitive statistical inferences about variances. *Organizational Behavior & Human Performance, 3*(2), 109-123.

Brunswick, E. (1956). *Perception and the Representative Design of Psychological Experiments*. University of California Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155 – 159.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85*(5), 395-416.

Engle, R. W. (2018). Working Memory and Executive Attention: A Revisit. *Perspectives on Psychological Science 13*(2), 190 – 193.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Hansson, P., Juslin, P., and Winman, A. (2008). The Role of Short-Term Memory Capacity and Task Experience for Overconfidence in Judgment under Uncertainty. *Journal Of Experimental Psychology: Learning, Memory, And Cognition, 34*(5), 1027-1042.

Henrion, M., & Fischhoff, B. (2002). Assessing uncertainty in physical constants. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment; heuristics and biases: The psychology of intuitive judgment.* Cambridge University Press, New York, NY.

Herdener, N., Clegg, B. A., Wickens, C. D., & Smith, C. A. P. (2019a). Anchoring and Adjustment in Uncertain Spatial Trajectory Prediction. *Human Factors, 61*(2), 255–272.

Herdener, N., Wickens, C. D., Clegg, B. A., & Smith, C. A. P. (2016). Overconfidence in projecting uncertain spatial trajectories. *Human Factors, 58*(6), 899 – 914.

Herdener, N., Wickens, C. D., Clegg, B. A., & Smith, C. A. P. (2017). Spatial Anchoring and Adjustment Under Mental Workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*(1), 349–349.

Herdener, N., Wickens, C. D., Clegg, B. A., & Smith, C. A. P. (2018). Attention Does Not Improve Impaired Understanding Of Variability In Spatial Prediction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 232–236.

Herdener, N., Wickens, C. D., Clegg, B. A., & Smith, C. A. P. (2019b). Can You See It? Perceived Variance in Scatterplot Visualization. To be presented at *The 63rd Human Factors and Ergonomics Society Annual Meeting.*

Hofstatter, P. R. (1939) Uber die Schatzung von gruppeneigenschaften. *Zeitschrift fur Psychologie, 145*, 1-44.

Horn, R. A. (2006). Understanding the repeated measures ANOVA. *Northern Arizona University*. Retrieved from http://oak.ucc.nau.edu/rh232/courses/eps625-phx/.

IBM. (2017). IBM Statistical Package for the Social Sciences [SPSS] Statistics for Mac, Version 25.0. Armonk, NY: IBM Corp.

Isadore, C. (2016). Kentucky Derby by the numbers. *CNN Business*. Retrieved from https://money.cnn.com/2016/05/05/news/companies/kentucky-derby-by-the-numbers/index.html.

Kahneman, D. (2011). *Thinking, fast and slow.* NY: Farrar, Straus, Giroux.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237 – 251.

Kareev, Y., Arnon, S., and Horwitz-Zeliger, R. (2002). On the misperception of variability. *Journal of Experimental Psychology, 131*(2), 287.

Lathrop, R. G. (1967). Perceived variability. *Journal of Experimental Psychology, 73*(4), 498-502.

Matthews, G., Davies, D. R., Westerman, S. J., & Stammers, R. B. (2013). Individual differences in ability and performance. In Matthews, G., Davies, D. R., Westerman, S. J., & Stammers, R. B. (Eds.), *Human Performance: Cognition, Stress and Individual Differences* (241 – 265). Psychology Press.

National Oceanic and Atmospheric Administration [NOAA]. (2017). Atlantic Basin Tropical Storms and Hurricanes: Track Errors (1970 - 2018). *National Hurricane Center Forecast Verification.* Retrieved from https://www.nhc.noaa.gov/verification/verify5.shtml.

National Oceanic and Atmospheric Administration [NOAA]. (2019). Definition of the NHC Track Forecast Cone. Retrieved from https://www.nhc.noaa.gov/verification/verify5.shtml.

Norwich, K.H. (1987). On the Theory of Weber Fractions. *Perception & Psychophysics 42*(3), 286 – 298.

Pallier, G., Wilkinson, R., Danthir, V., Kleitman, S., Knezevic, G., Stankov, L., and Roberts, R. D. (2002). The role of individual difference in the accuracy of confidence judgments. Journal of *General Psychology, 129*(3), 257-299.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*(1), 29-46.

Pollard, P. (1984). Intuitive judgments of proportions, means, and variances: A review. *Current Psychological Research & Reviews, 3*(1), 5 – 18.

Pugh, A. J., Wickens, C. D., Herdener, N., Clegg, B. A., Smith, C.A.P. (2018). Effect of visualization training on uncertain spatial trajectory predictions. *Human Factors, 60*(3), 324 – 339.

Qualtrics Software. (2019). Copyright 2019. Qualtrics and all other Qualtrics product or service names are registered trademarks or trademarks of Qualtrics, Provo, UT, USA. https://www.qualtrics.com.

R Studio Team (2015). RStudio: Integrated Development for R, Version 1.2.1335-1. RStudio, Inc., Boston, MA. http://www.rstudio.com/.

Rinne, L. F. & Mazzocco, M. M. (2013). Inferring uncertainty from interval estimates: Effects of alpha level and numeracy. *Judgment and Decision Making, 8*(3), 330 – 344.

Roediger, H. L. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology, 59*, 225 – 254.

Shiffrin, R. M., & Atkinson, R. C. (1969). Storage and retrieval processes in long-term memory. *Psychological Review, 76*(2), 179-193.

Smith, D. J., Zakrzewski, A. C., Johnson, J. M., Valleau, J. C., & Church, B. A. (2016). Categorization: The view from animal cognition. *Behavioral Sciences, 6*(12). doi:10.3390/bs6020012.

Spahr, K. S., Wickens, C. D., Clegg, B. A., Smith, C. A. P., & Williams, A. S. (2018). Calibrating Uncertainty: Commonalities in the Estimation of Numeric Variability Versus Spatial Prediction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 757–761.

Spencer, J. (1961). Estimating averages. *Ergonomics, 4*, 317-328.

Sternberg, R. (1999). *The Nature of Cognition*. MIT Press.

Tosto, M. G., Petrill, S. A., Halberda, J., Trzaskowski, M., Tikhomirova, T. N., Bogdanova, O. Y., Ly, R., Wilmer, J. B., Naiman, D. Q., Germine, L., Plomin, R, & Kovas, Y. (2014). Why do we differ in number sense? Evidence from a genetically sensitive investigation. *Intelligence*, *43*(100), 35–46.

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

Tyszka, T. & Zielonka, P. (2002). Expert Judgments: Financial Analysts Versus Weather Forecasters. *The Journal of Psychology and Financial Markets, 3*(3), 152 – 160.

University of Cambridge. (2018). Rules of thumb on magnitudes of effect sizes. *Cognition and Brain Sciences Unit.* Retrieved from http://imaging.mrc-cbu.cam.ac.uk/statswiki/ FAQ/effectSize.

Wickens, C.D., Clegg, B.A., Witt, J.K., Smith, C.A.P., Herdener, N., & Spahr, K.S. (2020). Model of variability estimation: factors influencing human prediction and estimation of variability in continuous information. *Theoretical Issues in Ergonomics Science, 21*(2), 220-238. DOI: 10.1080/1463922X.2019.1679907.

Wickens, C. D., Hollands, J. G., Banbury, S. & Parasuraman, R. (2013). *Engineering psychology and human performance* (4th ed.). New York: Pearson.

APPENDIX A

A power analysis was performed to estimate a proper sample size for the follow-up study, based on the correlation findings in Experiment 1[4] (*r* = 0.23, *p* = .21). This correlation is considered a moderate effect size using Cohen's (1992) guidelines. With a conservative ES = 0.20, alpha = .05, and power = 0.80, the minimum sample size required was 193 participants (Faul, Erdfelder, Buchner, & Lang [G*Power], 2009). This was proposed as part of the thesis, but such a large sample size comes with drawbacks, particularly when using Amazon Mechanical Turk (MTurk). In Experiment 1, participants were paid between $6.00 and $10.00, based on performance. With the minimum-powered sample size of 193 participants, the follow-up experiment would cost between $1,158.00 and $1,930.00 in compensation.

Due to these issues, the committee suggested a change in perspective and reduced scope for the remainder of this project. Specifically, this meant to just counterbalance the two tasks. This is an important next step, because if participants show a different pattern of results when the tasks are counterbalanced, this means that the order of the tasks influenced performance. This change in focus still answers the original research question (i.e., how do people understand mean and variability in spatial versus numeric stimuli?).

[4] The driving focus of this project was whether participants would show the same patterns of performance when responding to numeric and spatial variability (see Experiment 1, Hypothesis 1A). Results from Experiment 1 suggested a positive, moderate correlation (*r* = 0.23, *p* = .21) between numeric and spatial variability (Phase 1) estimates. Because of the importance of this correlation, this correlation coefficient was used as the parameter for the power analysis for the original Experiment 2.