

**UNDERSTANDING THE MISUNDERSTANDING: WHY CONFIDENCE  
INTERVALS ARE POORLY UNDERSTOOD AND EVALUATING  
PROPOSED SOLUTIONS ACROSS SENSORY MODALITIES**

A Dissertation  
Presented to  
The Academic Faculty

by

Jared M. Batterman

In Partial Fulfillment  
of the Requirements for the Degree  
Ph.D. in the  
School of Psychology

Georgia Institute of Technology  
August 2019

**COPYRIGHT © 2019 BY JARED M. BATTERMAN**

**UNDERSTANDING THE MISUNDERSTANDING: WHY CONFIDENCE  
INTERVALS ARE POORLY UNDERSTOOD AND EVALUATING  
PROPOSED SOLUTIONS ACROSS SENSORY MODALITIES**

Approved by:

Dr. Bruce N. Walker, Advisor  
School of Psychology  
*Georgia Institute of Technology*

Dr. John T. Stasko  
School of Interactive Computing  
*Georgia Institute of Technology*

Dr. Richard Catrambone  
School of Psychology  
*Georgia Institute of Technology*

Dr. S. Camille Peres  
School of Public Health  
*Texas A & M*

Dr. Jamie C. Gorman  
School of Psychology  
*Georgia Institute of Technology*

Date Approved: April 29, 2019

## ACKNOWLEDGEMENTS

This section is tough to write because there are too many people to thank and not nearly enough space, but I'm going to do my best to fit in as many people as possible. First and foremost, I want to thank my family. My wife Audrey, who was by my side throughout this process and drove me to cross the finish line. My kids Anna and Alex, who are the light of my life and reminded me why I was doing this to begin with. My father Scott and grandfather Steven who inspired me to become a rare third generation Ph.D. My mom Caren and siblings Alexa and Ethan who were always there to cheer me on and listen to me complain. Finally, all of my extended family who had my back whenever I needed it.

I'd also like to thank my Georgia Tech family: John, Jonathan, Alyssa, Brianna, Brittney, Keenan, Vincent, Synge, Rachael, and everyone else who helped me out and made a positive impact on my life. I'm proud to call all of you friends and I hope we stay that way in the years to come. I'd also like to thank the Nguyen family, whose help, care, and support for me, my wife, and my children helped enormously and made us feel like we had family nearby, even when everyone else was hundreds of miles away. Finally, I'd like to thank my advisor Bruce Walker for taking a chance on me and accepting me into his lab. The lessons I've learned there helped shape both my career and who I am as a person, and for that I'll always be grateful.

# TABLE OF CONTENTS

<b>ACKNOWLEDGEMENTS</b>	<b>iii</b>
<b>LIST OF TABLES</b>	<b>viii</b>
<b>LIST OF FIGURES</b>	<b>ix</b>
<b>SUMMARY</b>	<b>xi</b>
<b>CHAPTER 1. Introduction</b>	<b>1</b>
<b>1.1 Error Bars</b>	<b>2</b>
<b>CHAPTER 2. Education</b>	<b>8</b>
<b>CHAPTER 3. Visual Uncertainty Representations</b>	<b>17</b>
<b>3.1 Lessons for Error Bars</b>	<b>25</b>
<b>CHAPTER 4. Auditory Representations of Mathematical Data</b>	<b>29</b>
<b>4.1 ‘Standard’ auditory graphs: the basic architecture</b>	<b>29</b>
<b>4.2 The ‘Rules’: What works and what does not</b>	<b>30</b>
<b>4.3 The conceptual model of auditory graphs</b>	<b>32</b>
4.3.1 The task component	33
4.3.1.1 Nominal	33
4.3.1.2 Ordinal	34
4.3.1.3 Interval	34
4.3.1.4 Theoretical Predictions	35
<b>4.4 Listener Characteristics</b>	<b>35</b>
4.4.1 Commonalities	35
4.4.2 Individual Differences	38

4.4.3 Training and Learning	40
<b>4.5 Display Characteristics</b>	<b>41</b>
4.5.1 Data	42
4.5.2 Mappings, scalings, and polarities	42
4.5.3 Context	47
4.5.4 Temporal Characteristics	48
4.5.5 Multiple Data Series	49
<b>4.6 Interactions and mutual influences of task, listener, and display characteristics</b>	<b>50</b>
<b>4.7 Environmental concerns</b>	<b>50</b>
<b>4.8. A brief commentary on split attention and the redundancy paradox</b>	<b>51</b>
<b>CHAPTER 5. AUDITORY UNCERTAINTY</b>	<b>52</b>
<b>5.1 Introduction</b>	<b>52</b>
<b>5.2 Building from the ground up</b>	<b>52</b>
<b>CHAPTER 6. RESEARCH QUESTIONS</b>	<b>56</b>
<b>CHAPTER 7. METHODS</b>	<b>57</b>
<b>7.1. Study 1</b>	<b>57</b>
7.1.1. Participants	57
7.1.2. Experiment structure	57
7.1.3. Question selection	58
7.1.4. Alternative Display Creation	58
7.1.5. Procedure	60
7.1.6. Hypotheses/Expected Results	60
<b>7.2. Study 2</b>	<b>61</b>

7.2.1. Participants	62
7.2.2. Experiment structure	62
7.2.3. Question selection	63
7.2.4 Auditory Stimulus Creation	63
7.2.5 Supplemental Materials Creation	64
7.2.6 Procedure	65
7.2.7. Hypotheses/Expected Results	65
<b>CHAPTER 8. Data Analysis and Results</b>	<b>67</b>
<b>8.1. Study 1 Analysis</b>	<b>67</b>
8.1.1. Study 1 Results	67
8.1.2. Study 1 Data Explained	70
8.1.3. Study 1 Summary Discussion	72
<b>8.2 Study 2 Analysis</b>	<b>73</b>
8.2.1. Study 2 Results	74
8.2.2. Study 2 Data Explained	76
8.2.3. Study 2 Summary Discussion	78
<b>CHAPTER 9. GENERAL DISCUSSION AND CONCLUSIONS</b>	<b>80</b>
<b>APPENDIX A: QUESTIONS FOR EXPERIMENTS</b>	<b>83</b>
Appendix A1. Confidence Interval Questions with No Graphs	83
Appendix A2. Confidence Interval Questions with Graphs	87
Appendix A3. Confidence Interval Questions with Error Bars	91
Appendix A4. Confidence Interval Questions with Alternative Representation	95

Appendix A5. Probability Questions with No Graphs	99
Appendix A6. Probability Questions with Graphs	101
<b>APPENDIX B: TRAINING SCRIPTS</b>	<b>104</b>
Appendix B1. Confidence Interval Training Script	104
Appendix B2. Probability Training Script	107
<b>APPENDIX C: SUPPLEMENTAL MATERIALS</b>	<b>109</b>
Appendix C1. MathML Code	109
Appendix C2. Accessible Statistics Tables	132
<b>REFERENCES</b>	<b>159</b>

## LIST OF TABLES

Table 1 - A list of relevant educational standards.	7
Table 2 - a. depicts the effect of visual aids on confidence interval questions versus those on another topic, b. splits the data further to look at the effects of data tables and graphs on these question types, c. depicts the differences between responses on the first five standard questions on the exam versus the investigative task across these groups, and d. depicts the same differences within question types.	12
Table 3 - A depiction of the glyph types tested by Wittenbrink et al. (1996). The first column contains the glyphs themselves and the second and third columns (labelled $d\theta$ and $dm$ ) contain xs or blank spaces indicating whether or not they are intended to represent uncertainty using an angle and/or a magnitude (size).	18

## LIST OF FIGURES

- Figure 1 - An example of a single trial in Belia et al.'s (2005) experimental task. Trials varied on key components (what error bars represent, ns of both groups, types of statistical test performed, etc.), but the primary task (moving the error bars to point of significance) stayed the same. 4
- Figure 2 - Two data points (labelled A and B) surrounded by error bars. 6
- Figure 3 - An example of the wind velocity display used by Wittenbrink et al. (1996). The display on the left shows wind velocity (speed and direction) without showing uncertainty. The display on the right uses the same data, but utilizes exaggerated angles and variations in arrow width to convey uncertainty. 20
- Figure 4 - The image on the left is a still from Andre and Cutler's (1998) first experiment, depicting the operator's ship, the meteor (with uncertainty), and its projected path. Though the actual experimental task was in color, only black and white pictures were provided. The image on the right is my coloration of the uncertainty displays to better represent what they actually looked like in the study. 21
- Figure 5 - The left image is a still from Andre and Cutler's (1998) second experiment depicting the gun turret at the bottom of the screen and the friendly and enemy ships flying above it. The right image depicts the uncertainty displays used in the study: the numeric, the graphical ring, and the graphical arc. 22
- Figure 6 - The left image depicts the beginning of a trial in Finger and Bisantz's (2002) decision making task. The right image depicts the three types of icons participants will be exposed to and how they are classified by the experimenters. 24
- Figure 7 - The representations used by Padilla et al. (2015) to represent uncertainty. The leftmost representation is a standard error bar, while the other representations were created for the purposes of their studies to evaluate their utility. 25
- Figure 8 - Sample visual graph to be sonified. Auditory equivalent explained in text with sound file available upon request. 30
- Figure 9 - Image A is an example of the gestalt principle of proximity. The dots on the left portion of the image are very close to each other so they are perceived as a single object. The dots on the right portion are further apart so they are perceived in 3 separate groups. Image B shows two shapes with visible portions missing that through the principle of continuation are viewed as single contiguous objects. 37
- Figure 10 - An image of a trial from Walker's (2002) conceptual magnitude estimation study. The conceptual dimension (in this case temperature) is on the top left, the button in the center plays the audio, the response box in the bottom right is left blank for the participant to assign a number to the auditory stimulus, and the next button moves them on to the next trial. 43
- Figure 11 - An image of a trial from Walker's (2002) follow up study assessing the efficacy of conceptual magnitude estimation in a more applied context. Clicking the 'START HERE' button will play the series of sounds and then participants choose which numeric range they feel best represents the concept of size by clicking the 'Scenario A' or 'Scenario B' buttons. 45

Figure 12 - Bar graph depicting differences in scores between question types collapsed across training condition. Error bars represent 0.95 confidence intervals.	68
Figure 13 - Bar graph depicting differences in scores between split by both question type and training condition. Error bars represent 0.95 confidence intervals.	69
Figure 14 - Bar graph depicting differences in scores between training conditions for both 1st and 2nd order filters. Error bars represent 0.95 confidence intervals.	74
Figure 15 - Bar graph depicting differences in scores between conditions for participants who identified as sighted, blind, and visually impaired. Error bars represent 0.95 confidence intervals.	76
Figure 16 - Bar graph depicting differences in scores between conditions for participants who identified as sighted vs. those who did not. Error bars represent 0.95 confidence intervals.	77

## SUMMARY

The error bar representation of a confidence interval is the most ubiquitous display of uncertainty in statistical analysis. However, despite this, error bars are poorly understood even by seasoned scientists and researchers across disciplines. The root of this misunderstanding is not certain, but researchers have posited several hypotheses ranging from the structure of the display itself, to how it is presented in the classroom.

Studies have thus far been either incomplete or inconclusive, leading some to call for the elimination of the use of error bars entirely. However, research into statistics education (suggesting the error bar representation may not even be taught in contemporary classrooms) demonstrates that a lack of exposure in an educational likely contributes to the displays' poor level of discernment. Furthermore, promising research into auditory uncertainty displays suggests that sound may be a better medium to present this information than vision.

The studies conducted in this dissertation demonstrate that not only are auditory displays viable alternatives to visual representations of confidence intervals (i.e., error bars), but when used together can improve overall performance. Furthermore, the presence of visual displays was shown to enhance performance on questions involving confidence intervals, but not on other topics, suggesting that confidence intervals may hold a unique status when it comes how they are affected by alternative display types.

Limited training was also shown to improve performance, suggesting that if confidence interval displays were taught in classrooms, overall performance could improve and lead students to understand this topic just as well as other statistical concepts. The success of the

auditory representation could also lead to the creation of cheaper, more universally designed, materials to help convey uncertainty to all students, regardless of sensory ability.

## CHAPTER 1. INTRODUCTION

In order to parse through, and make sense of, the immense amount of data presented in modern society, statistical literacy is essential. Unfortunately, understanding of key statistical concepts remains poor. Furthermore, this lack of understanding is not limited to the general populace, and even extends to scientific practitioners who rely on statistical techniques for their work. One of the most commonly misunderstood of these often-deployed techniques is the confidence interval (Belia, Fidler, Williams, & Cumming, 2005; Cumming, 2009; Cumming, Fidler, & Vaux, 2007; Cumming & Finch, 2001, 2005; Cumming & Maillardet, 2006; Cumming, Williams, & Fidler, 2004; Fidler & Cumming, 2005; Fidler, Thomason, Cumming, Finch, & Leeman, 2004; Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Kalinowski, 2010).

Introduced in the late 1930s, the confidence interval has seen continued use in all facets of science as a reliable and concrete way to display the uncertainty associated with a given data point (Belia et al., 2005; Hoekstra et al., 2014; Neyman, 1937). However, in recent years the use of confidence intervals has exploded, with many arguing that the confidence interval should be used in lieu of the more common Null Hypothesis Significance Testing (NHST). Though the argument against NHST began to ramp up in the 1960s, the fever pitch that it has reached in recent years has led some journals to attempt to limit its use in their publications (with one going so far as to ban it) (Bakan, 1966; Cortina & Dunlap, 1997; Leek & Peng, 2015; Rozeboom, 1960; Woolston, 2015). Regardless of one's views on NHST, its fall from favor has pushed confidence intervals to the forefront as one of the premiere ways for researchers to present their data. However, this has led to the realization that this increased use has not led to increased understanding, even amongst the most learned practitioners.

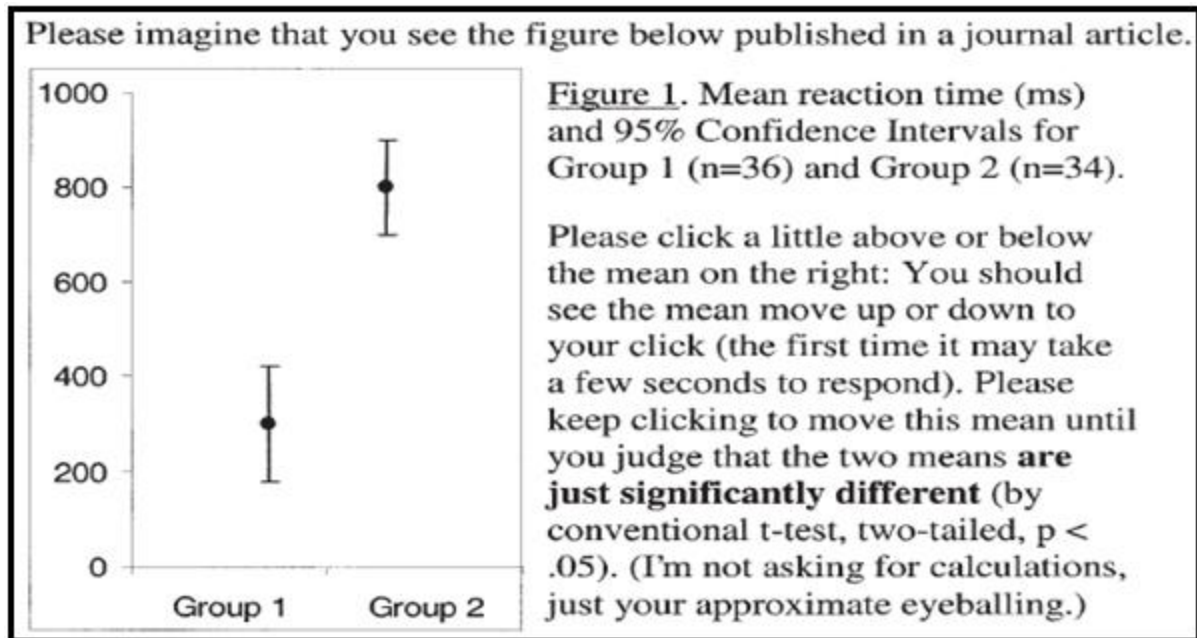
## 1.1 Error Bars

In scientific publications, confidence intervals are most often displayed visually using error bars (Krzywinski & Altman, 2013; Payton, Greenstone, & Schenker, 2003; Student, 1908). Error bars take the form of single lines intersecting a data point and are often capped with smaller perpendicular brackets. They most commonly represent either the standard error of the mean or the boundaries of a confidence interval (Field, 2009; Tukey, 1977). Though these are statistically disparate concepts, they share the same representation. The only way to know which of these concepts a set of error bars represents is to be informed by the author. In addition to encapsulating the relative uncertainty attached to a given data point, error bars are also meant to be used as a method of statistical comparison. Their relative degree of overlap indicates whether or not a given data point is statistically different from another data point. This is extremely important because statistical difference is oftentimes the aim of scientific studies, with error bars being the visualization of choice for nearly all practitioners. Though error bars are ubiquitous in academic publications (with many journals requiring them), research has demonstrated that they are not very well understood, even by seasoned researchers (Belia et al., 2005; Cumming, 2009; Cumming et al., 2007; Cumming & Finch, 2001, 2005; Cumming & Maillardet, 2006; Fidler, Cumming, et al., 2005; Fidler et al., 2004; Fidler, Thomason, Cumming, Finch, & Leeman, 2005).

Though error bars have been a point of contention in science for a while, it is only in the past couple decades that the debate over their understanding and use has truly reached its zenith (Belia et al., 2005; Correll & Gleicher, 2013, 2014; Krzywinski & Altman, 2013). The current debate was kicked off by Dr. Geoff Cumming and his research group out of La Trobe University. Their first study in this line of research (Cumming, Williams, & Fidler, 2004) consisted of

surveying research professionals in three fields (psychology, behavioral neuroscience, and medicine) to determine how well they understood confidence intervals (as represented visually by error bars) in terms of their relation to a study's repeatability. To do this, they presented their participants with various means and confidence intervals and tasked them with estimating plausible replication means. Though this task may not have ecological validity compared to actual error bar interpretation, the authors wanted to evaluate the practitioners' implicit understanding of the concept, and create a scenario where they would not attempt to calculate a numerical answer before responding. Results demonstrated no difference between the three fields, but what seems to be either a misunderstanding of the variability of replication means or a somewhat high expectation for what should qualify as a valid or successful replication. The majority of responders offered replication means that suggested that they believed the means would almost always fall within the bounds of a 95% confidence interval and only deviate from this at the colloquial rate of 5%. However, in reality, only about 5 out of 6 replication attempts will fall into this range, leading to a more accurate failure rate of around 17% percent (Estes, 1997; Tversky & Kahneman, 1971). Though the authors admit that this failing could likely have more to do with statistics education than display, they also argue that a well-constructed display should convey this information in an easier to understand manner.

This study was followed up by a much more thorough investigation of error bars and their interpretation (Belia et al., 2005). As opposed to the more indirect route taken by their previous study, this study focused on asking researchers direct questions about error bars and their interpretation. Furthermore, participants in this study were tasked with not only interpreting standard error bars and confidence intervals, but also manipulating them to determine whether or



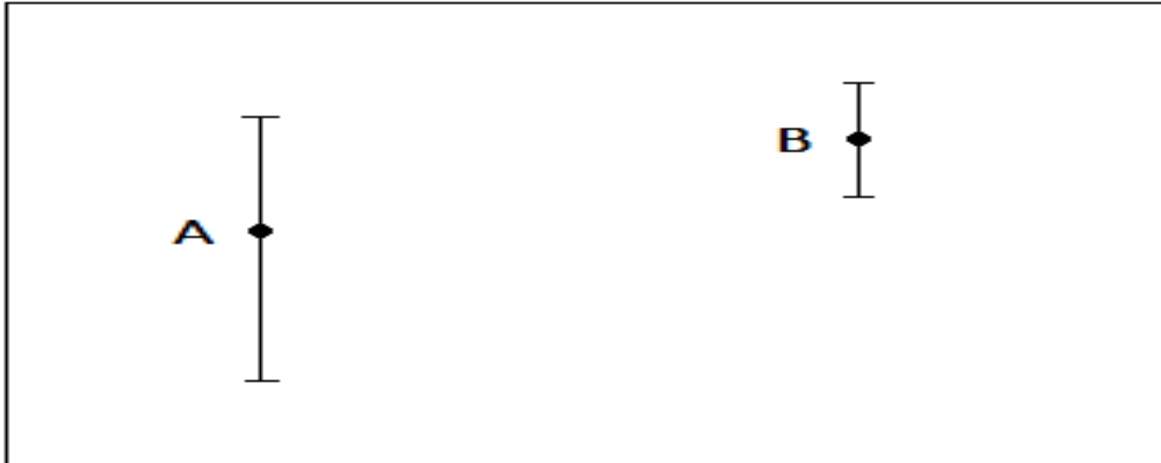
**Figure 1 - An example of a single trial in Belia et al.'s (2005) experimental task. Trials varied on key components (what error bars represent, ns of both groups, types of statistical test performed, etc.), but the primary task (moving the error bars to point of significance) stayed the same.**

not they understood the subtleties of how the bars themselves related to statistical significance. Like their prior study, their participants were all seasoned researchers (with multiple refereed publications in top journals) from the fields of psychology, behavioral neuroscience, and medicine. The experimental task consisted of participants being presented with a pair of data points surrounded by error bars of various sizes, given information about the data and the tests performed on it, and asked to move the error bars to the point where they were just statistically significant (see Figure 1). Results revealed a troubling misunderstanding of both how error bars relate to statistical significance, and the interpretational differences between standard error bars

and confidence intervals. Furthermore, research experience, years in the field, and publication count had no effect on participants understanding.

The authors broke down their findings into four different problems/misconceptions. First, participants were not only inaccurate, but oftentimes extremely inaccurate, with only 22% of participants moving the bars to correspond with p-values between 0.025 and 0.10. Second, participants did not distinguish between standard error bars and confidence intervals at all, as their responses were near identical for both display types. Third, 31% of respondents adhered to the incorrect rule of thumb that when error bars (regardless of whether they are meant to represent standard error or the bounds of confidence intervals) are just touching, the means are statistically different at a  $p < 0.05$  level. Fourth, when given information about the experimental design (i.e. whether it was between-subjects or within-subjects/repeated measures), participants either did not understand the effect this had on interval interpretation or ignored it entirely, as providing the information had no effect on responses.

Cumming and colleagues followed up this study with other evaluations, but they all came to very similar conclusions: practitioners (regardless of field or experience) do not understand how error bars relate to significance, do not understand the difference between standard error bars and confidence intervals, and they do not take key factors of experimental design (such as whether the study/analysis was conducted within or between-subjects) into consideration when evaluating these intervals (Belia et al., 2005; Beyth-Marom, Fidler, & Cumming, 2008; Coulson, Healey, Fidler, & Cumming, 2010; Cumming, 2009). Though there have been many proposed solutions to these problems, they tend towards the vague (i.e. we need more/better statistical education) or the extreme (we need to do away with error bars/uncertainty representations entirely)



**Figure 2 - Two data points (labelled A and B) surrounded by error bars.**

(Beyth-Marom et al., 2008; Correll & Gleicher, 2013, 2014; Fidler & Cumming, 2007; Fidler et al., 2004; Fidler, Thomason, et al., 2005; Scown, Bartlett, & McCarley, 2014). It is almost as if the research community has decided that the solution to this problem is binary, get rid of the current representation or keep it and try to somehow reclaim it in with the next generation of researchers, who will be presumably be somehow better educated in their use.

Getting rid of error bars entirely seems unnecessary and counterproductive; especially because though their use as a tool to convey statistical difference may be suspect, they serve other purposes that are worth mentioning. Specifically, for experienced researchers, error bars are effective at conveying the uncertainty inherent in a given data point and allow for basic size comparisons. For example, Figure 2 depicts two data points absent context. Though nothing is known about these data, basic familiarity with error bars allows us to determine that point A has a given uncertainty value, point B has a given uncertainty value, and the value associated with point A is greater than the one associated with point B. All of that said, as a medium for communicating higher level uncertainty (especially to a statistical degree of precision), error bars

**Table 1 - A list of relevant educational standards.**

Common Core Mathematics Standards:	
ID Number	Text of Standard
HS.S-IC.4.	Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling.
HS.S-IC.5.	Use data from a randomized experiment to compare two treatments; use simulations to decide if differences between parameters are significant.
Georgia State Mathematics Standards:	
ID Number	Text of Standard
MGSE9-12.S.IC.1	Understand statistics as a process for making inferences about population parameters based on a random sample from that population.
MGSE9-12.S.IC.4	Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling.
MGSE9-12.S.IC.5	Use data from a randomized experiment to compare two treatments; use simulations to decide if differences between parameters are significant.
MGSE9-12.S.IC.6	Evaluate reports based on data. For example, determining quantitative or categorical data; collection methods; biases or flaws in data.
Georgia State Statistical Reasoning Standards:	
MSR.IR.1	Students will ask if the difference between two sample proportions or two sample means is due to random variation or if the difference is significant. Students will be able to determine if there are differences between two population parameters or treatment effects.

are not ideal. Furthermore, having them tasked with representing two similar but disparate concepts (standard error and confidence intervals) is demonstrably too much for the current iteration of the display to handle. There are two main ways to address this problem: first by looking at education, and second by taking the lessons of more successful visual uncertainty displays to heart and incorporating them into the existing display.

## CHAPTER 2. EDUCATION

To understand why confidence intervals are so misunderstood, it is prudent to look at how they are currently taught. The best starting point for this is to investigate the current educational standards that pertain to them. The most pertinent mathematics standards that have the furthest reach nationally are the Common Core ("Mathematics Standards," 2017). On a more local level, the state of Georgia also adheres to their own set of general mathematics standards in addition to a separate set of standards pertaining solely to statistics ("Georgia Department of Education Statistical Reasoning Standards," 2017; "Mathematics Georgia Standards of Excellence," 2017). All sets of standards were thoroughly investigated to determine which most pertained to the study and understanding of confidence intervals (see Table 1).

Unfortunately, though some of the standards pertain to similar concepts to those conveyed by confidence intervals (statistical significance, general inference, and differences between sample and population parameters) none of them mention them directly. In addition, none of them focus on interpretation of graphical displays. As investigating standards in lower (K-12) education did not yield much useful information, my focus shifted towards college level statistics courses. Unsurprisingly, there is tremendous variation in how concepts are taught, the order they are presented in, and even what knowledge a student is expected to have upon completion of the course (Batanero, Burrill, & Reading, 2011; Blalock Jr, 1987; Garfield, 1993; Horton & Hardin, 2015; Moore, 1988; Ricketts & Berry, 1994; Schmidt, McKnight, & Raizen, 2007). However, there is one agreed upon assessment that nearly every university in the United States seems to agree on: The Advanced Placement (AP) Statistics exam ("The AP Statistics Exam," 2017). A score of 3 or higher (with 4s or 5s required at some top tier universities) is accepted by thousands

of universities across the country as equivalent to the completion of an introductory college level statistics course ("Statistics Course Description," 2010). This means that it is logical to assume that the contents of this exam reflect the base knowledge universities expect of students after a semester of statistics. Like with every AP exam, the College Board provides a plethora of preparatory materials including practice tests, lesson plans for courses, and the broader conceptual themes that are most important for a student to understand. For their statistics course, they break down their content into four key areas of focus that represent percentages of their overall course goals (as well as the percentage of the exam that focuses on each area):

- 1) Exploring Data: Describing patterns and departures from patterns (20-30%)
- 2) Sampling and Experimentation: Planning and conducting a study (10-15%)
- 3) Anticipating Patterns: Exploring random phenomena using probability and simulation (20-30%)
- 4) Statistical Inference: Estimating population parameters and testing hypotheses (30-40%)

In turn, each of these course goals is broken down into subsections of information that students are expected to know in order to perform well on the AP exam. Confidence intervals make up half of the statistical inference section under the subheading 'Estimation (point estimators and confidence intervals)' (see full text below).

“A. Estimation (point estimators and confidence intervals)

1. Estimating population parameters and margins of error
2. Properties of point estimators, including unbiasedness and variability
3. Logic of confidence intervals, meaning of confidence level and confidence intervals, and properties of confidence intervals
4. Large sample confidence interval for a proportion

5. Large sample confidence interval for a difference between two proportions
6. Confidence interval for a mean
7. Confidence interval for a difference between two means (unpaired and paired)
8. Confidence interval for the slope of a least-squares regression line”

However, despite the prominent role of confidence intervals in the course curriculum, visual representations of them do not appear at all in any of the provided preparatory materials ("Statistics Course Description," 2010). Furthermore, these representations also do not appear in some of the most popular third party test guides (Hinders, 2015; Mulekar, 2015; Sternstein, 2013). Though initially puzzling, the reasons for this became clear: they never appear on the exam itself.

The AP exam itself is split into two sections; a multiple choice and a free response section. The multiple-choice section consists of 40 questions and is worth 50% of the total exam score; the free response section is worth the other 50% and consists of 6 questions (5 short answer questions and 1 investigative task). Though the actual multiple-choice portion of past exams is difficult to obtain (the publishers generally do not release them except in rare cases: see the 2012 and 1997 exams), every free response section has been made freely available since 2003. In addition to releasing these (free response) questions, the College Board also releases some de-identified student responses and full score reports which include the correct answers; as well as summary statistics of students' overall performance, grader commentary, and areas students excelled at or struggled with for each of the questions. These reports also include the topic area and goals each question attempts to assess. However, amongst all of these questions and reports, visual representations of confidence intervals themselves appear a total of zero

times. This is despite the fact that confidence interval questions were present 11 out of the 14 years (2003-2016) that score reports were released. Furthermore, of the 11 years that they were present on the exam, they performed worse than all other topics on 7 of them (i.e. 63.6% of the time). As a whole (including the years when confidence intervals questions scored higher than other topics), confidence interval questions performed worse overall (1.377 out of 4 versus 1.413 for others). In order to determine why this is the case, confidence interval questions were further explored to determine what (if any) aspects of them affected overall student performance.

The primary facet explored was the presence of visual aids. Though no direct representation of confidence intervals were present on any of the released free response questions, some of the questions did provide visual aids (which are operationally defined as either graphs or data tables). When scores are split in this manner, it reveals an interesting performance gap: questions with visual aids present had an average score of 1.43, while those without them had a score of 1.285. To determine whether the type of visual aid mattered, questions were compared based on whether they contained graphs or data tables. In this instance, questions with graphs present scored 1.54, while those with data tables present scored 1.333. However, this analysis included questions in which both graphs and data tables were present. When those questions are removed, the difference between these categories becomes even starker with graphs questions averaging 1.56 and data table questions averaging 1.155, while questions with both present averaged 1.51. This suggests a few things: first, that when it comes to confidence interval questions, graphs are helpful, graphs and data tables together are slightly less helpful, and data tables alone are potentially more detrimental than no visual aids at all (see Table 2a).

**Table 2 -**

**a.**

	CI (Mean)	NCI (Mean)	Cohen's d
All	1.377	1.413	-0.03
VA	1.430	1.484	-0.05
NVA	1.285	1.342	-0.05
Just DT	1.155	1.489	-0.30
Just Graphs	1.560	1.417	0.13
Both	1.510	1.519	0.00

**b.**

	Cohen's d
CI VA vs NVA	0.13
CI Graphs vs DT	0.2
CI Just Graphs vs Just DT	0.39
CI both vs Just Graphs	0.05
CI both vs Just DT	0.34
NCI VA vs NVA	0.13
NCI Graphs vs DT	-0.03
NCI Just Graphs vs Just DT	-0.06
NCI both vs Just Graphs	0.09
NCI both vs Just DT	0.03

**c.**

	CI	NCI	Cohen's d
1to5	1.367	1.437	-0.06
6	1.395	1.371	0.02
1to5VA	1.420	1.530	-0.10
1to5NVA	1.297	1.342	-0.04
1to5 Just DT	1.155	1.511	-0.32
1to5 Just Graphs	1.685	1.600	0.08
1to5both	x	1.520	x
6VA	1.443	1.361	0.07
6NVA	1.250	x	x

**d.**

	Cohen's d
CI 6 vs 1-5	0.03
CI 1-5VA vs NVA	0.12
CI 1-5 Just Graphs vs Just DT	0.52
CI 1-5 Both vs Just Graphs	x
CI 1-5 Both vs Just DT	x
CI 6VA vs 6NVA	0.18
NCI 6 vs 1-5	-0.06
NCI 1-5VA vs NVA	0.18
NCI 1-5 Just Graphs vs Just DT	0.08
NCI 1-5 Both vs Just Graphs	-0.07
NCI 1-5 Both vs Just DT	0.01
NCI 6VA vs 6NVA	x

**Table Notes - a.** depicts the effect of visual aids on confidence interval questions versus those on another topic, **b.** splits the data further to look at the effects of data tables and graphs on these question types, **c.** depicts the differences between responses on the first five standard questions on the exam versus the investigative task across these groups, and **d.** depicts the same differences within question types.

Though these descriptive statistics are useful, it would be ideal if more advanced statistical measures could be applied. Unfortunately, the nature of the data provided only allows for limited options. An analysis of variance is impossible because the score reports do not provide the variance of the difference scores (or any difference scores at all) and paired comparisons (i.e. t-tests) would be uninformative because each mean is derived from around 50,000-200,000 responses and any difference in any direction would appear significant regardless of whether or not the difference was truly there (i.e. a type I error would likely occur). However, despite the ineffectiveness of traditional significance testing on this dataset, effect sizes could easily be calculated to determine the magnitude of the differences between groups. Due to the nature of the data, Cohen's  $d$  is the best method to quantify this and was calculated for all relevant mean comparisons. These calculations corroborate what the means suggested, that graphs improve performance on confidence interval questions to a substantive degree, while data tables absent graphs detract from it (see Tables 2a-d for all means and effect sizes).

In addition to exploring questions on confidence intervals, all other topics were also explored to determine whether the pattern of results observed with confidence intervals were unique to them or held true for other subjects. Results revealed that the presence or absence of visual aids as a whole was just as impactful for other topics as they were for confidence intervals ( $d=0.13$ , see Table 2b); however, that is where the similarities stop. With confidence intervals, graphs were extremely helpful, while data tables detracted, and both together were overall helpful, but not as much as graphs alone. With the other topics, data tables alone outperformed graphs, and both together outperformed both graphs and data tables. This suggests that for other topics, graphs are not as overall helpful as they are for confidence interval questions. This is interesting not only because it suggests a unique status for these sorts of questions, but also

because the graphs shown on confidence interval questions never depict confidence intervals themselves.

The last analyses conducted split the traditional free response questions (1-5 on the exam) from the investigative task (question 6) to determine any overall performance differences. Results show near identical scores and a very similar pattern of results across both question types for both confidence interval and other topics questions. The only notable difference seems to be an even larger effect of graphs on confidence interval questions in questions 1-5 compared to the investigative task. However, it is likely that the reason for this is a lack of sufficient data on investigative tasks. Because the overall scores and result patterns between the investigative task and other questions are similar, it is likely this difference would be replicated with more data. But because each year provides a 5:1 ratio of standard free response questions to investigative tasks, comparable levels of data would be difficult to obtain short of the college board releasing decades more testing data.

In addition to thoroughly analyzing the quantitative data in the provided score reports, the qualitative data were also analyzed in order to determine whether the scorers noticed any patterns in results, had any helpful commentary, etc. For each question on each score report the data are presented in the same way. First it will list the intent of the question (i.e. what the question was supposed to assess, which topics, etc.), then the mean and standard deviation of the scores for that question, then common student errors and omissions, and then a section of advice for AP educators on how to present the material better in the future to improve scores. For confidence interval questions, the content of the score reports generally noted similar items. Students tended to make the same sorts of errors; such as accidentally calculating a z-interval, confusing confidence intervals and confidence levels, various misunderstandings of the central limit

theorem such as assuming a dataset with an  $n$  under 30 could not be normal, and many errors in interpretation stemming from a fundamental misunderstanding of what a confidence interval is and what it gauges. The advice for educators primarily focuses on addressing these challenges by stressing more frequent practice and emphasizing the differences between oft confused concepts. Though the score report content is normally systematic and similar, in some cases it differs in interesting ways revealing details about both the exam itself and confidence interval questions as a whole.

One of the most notable examples of this occurs in the errors section of the first free response question from the 2013 exam. This question provided data in a stem and leaf plot and asked test takers to use this data to identify, compute, and interpret a confidence interval based on it. Performance on this question was exceptional in comparison to other questions from other years on this topic, with an average score of 2.27 (the highest among confidence interval questions released from 2003-2016). Furthermore, it was the highest scoring free response question of that year (2013) regardless of topic. As to why the question scored so highly, no suggestion was given. However, the errors section revealed that students did something on this question that they did not do with other similar ones:

“A surprising number of students entered the data into a calculator and produced a graphical display in addition to the stem-and-leaf plot provided in the question. This was not an error, but was an unnecessary step that wasted valuable time.”

So, not only were test takers provided a graphical display (a stem and leaf plot), but they also created additional ones to help them answer the question. What is most surprising about this is







that the educators that graded the test viewed this in a negative manner, seeing it only as a waste of time. This further suggests that educators themselves are not aware of what may be contributing both to higher scores and better understanding.

### CHAPTER 3. VISUAL UNCERTAINTY REPRESENTATIONS

One of the continuing trends evident throughout the analyses of AP test questions is that the presence of visual aids (namely graphs for confidence interval questions) as a whole is beneficial and is reflected in overall scoring. However, it is important to take into account exactly how data are being represented. Though the AP exam neglects to present or assess students on their knowledge of visual confidence interval representations, it may also be the case that the ‘error bar’ representation itself may not be conducive to both student learning and practitioner understanding. Despite a massive amount of research into visual uncertainty representations, the error bar representation has hardly changed since its inception. Though limited education and exposure is certainly a part of the representations’ weak levels of understanding, its unknown whether its overall design could contribute to this.

The body of research suggests that the most effective uncertainty visualizations share qualities with what engineering/human factors psychologists refer to as configurable displays and computer vision and graphics researchers refer to as visual glyphs (Bisantz et al., 2011; Pang, Wittenbrink, & Lodha, 1997; Stokes, Wickens, & Kite, 1990; Wickens, 1992; Wickens & Carswell, 1995; Wittenbrink, Pang, & Lodha, 1996). These displays contain emergent features, defined as aspects of a display that emerge through an interaction of their graphical elements to provide easily perceptible information about a given system, variable, or set of variables (Bennett, Toms, & Woods, 1993; Bisantz et al., 2011; Finger & Bisantz, 2002). Though ubiquitous today, these displays trace their origins to early 20<sup>th</sup> century military interfaces,

**Table 3 -**

	$d\theta$	$dm$
	X	X
	X	
	X	X
		X
	X	X
	X	X

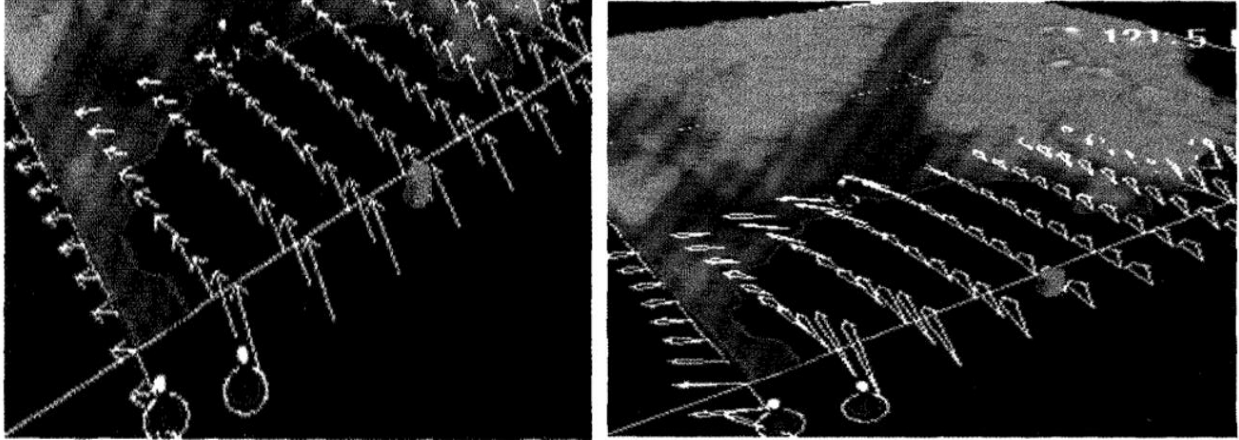
**Table Notes - A depiction of the glyph types tested by Wittenbrink et al. (1996). The first column contains the glyphs themselves and the second and third columns (labelled  $d\theta$  and  $dm$ ) contain xs or blank spaces indicating whether or not they are intended to represent uncertainty using an angle and/or a magnitude (size).**

notably polar coordinates displays which were investigated thoroughly in the 60s, 70s, and 80s [see (Coekin, 1970; Lee & Kirlik, 2013; Woods, Wise, & Hanes, 1981) for an in depth look at this research]. However, their use as functional uncertainty displays is much more recent.

Wittenbrink et al. (1996) investigated how to optimally represent uncertainty when displaying weather and environmental data in order to improve the accuracy of user interpretations. To do this, they designed a variety of visual glyphs to convey multiple types of uncertainty both individually and simultaneously. For example, when representing wind currents, they needed their glyphs to convey uncertainty about both wind speed and direction. However, in some cases users may only want/need one type of uncertainty available at a time; so, both of these circumstances needed to be accounted for. Because this sort of glyph had never

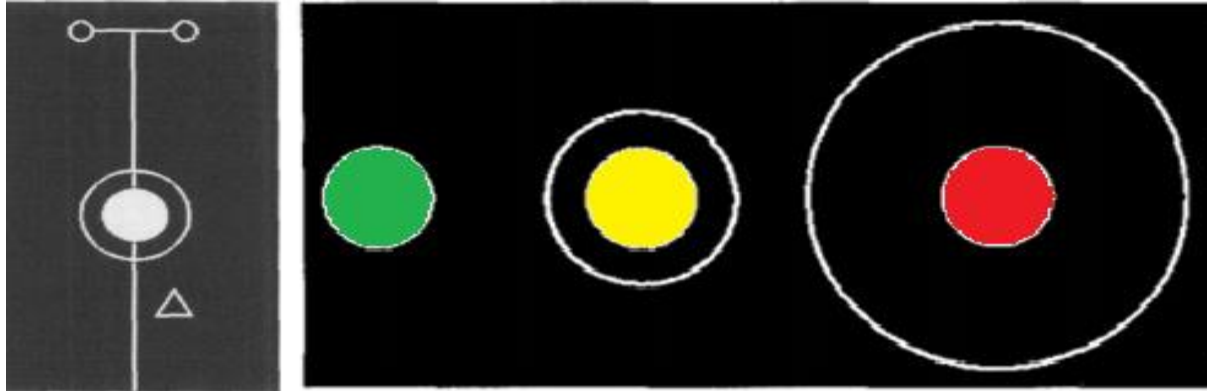
been tested or deployed in this context before, the authors wanted to test multiple glyph types to both compare their effectiveness and to get a better sense of what aspects of their visualizations worked and which did not (see Table 3).

To eliminate any potential confounds due to ignorance of the concepts being displayed, the authors chose to test their visualizations on highly skilled participants (individuals in the United States Military pursuing Masters degrees in meteorology or related sciences). Furthermore, to make the task itself more realistic they pulled real weather data from the Geographic Information Systems (GIS). Because the primary focus of these glyphs was to increase accuracy in interpreting environmental data, the authors were not concerned with task completion time. As such, their analyses focused primarily on the number of errors participants made in their interpretations of data and not how quickly they were able to complete a given task. Tasks consisted of a series of questions associated with a series of diagrams randomly assigned to represent the same information in different ways both with and without uncertainty present. Results demonstrated that though participants seemed to understand all of the uncertainty representations presented, they only significantly increased performance in one case. That said, their presence never decreased performance and at their worst, they performed equivalently to displays without them while also offering more information. Significant performance increases occurred when participants were asked to interpret and predict wind velocity using directional arrows. In the control (no uncertainty) condition, the length of the arrows represented wind speed (longer arrows meant faster winds) and their angle relevant to the horizontal plane indicated their



**Figure 3 - An example of the wind velocity display used by Wittenbrink et al. (1996). The display on the left shows wind velocity (speed and direction) without showing uncertainty. The display on the right uses the same data, but utilizes exaggerated angles and variations in arrow width to convey uncertainty.**

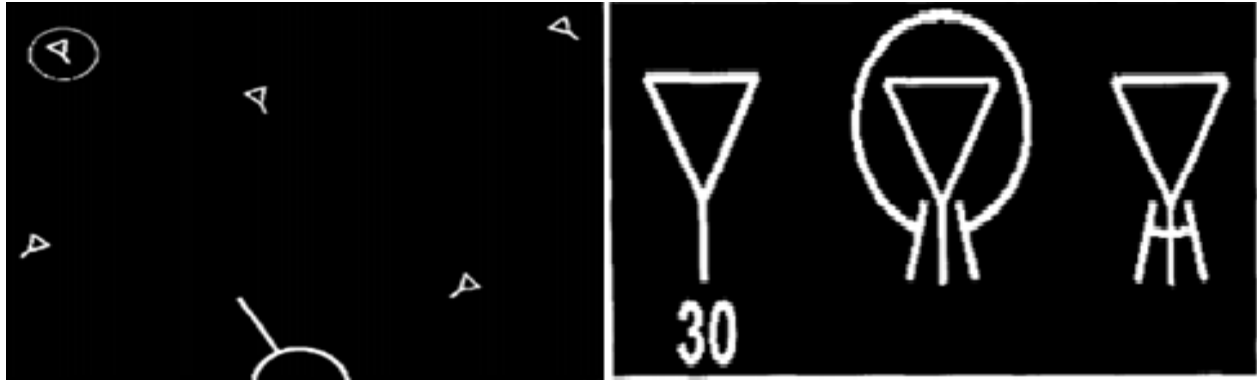
direction. In the uncertainty condition, the basic display was the same, but the width of the arrows was varied to represent the uncertainty associated with wind speed and their angle was exaggerated positively or negatively to indicate uncertainty associated with wind direction (see Figure 2). Participants consistently performed better in the uncertainty condition in every case, suggesting that the display was superior both at interpreting static values and at accurately predicting future ones. There are a few main takeaways from this study, but first and foremost it demonstrated that uncertainty can be presented in tandem with the value(s) it



**Figure 4 - The image on the left is a still from Andre and Cutler's (1998) first experiment, depicting the operator's ship, the meteor (with uncertainty), and its projected path. Though the actual experimental task was in color, only black and white pictures were provided. The image on the right is my coloration of the uncertainty displays to better represent what they actually looked like in the study.**

is in reference to (i.e. the level of uncertainty associated with wind speed, etc.) without detracting from performance, and at times enhancing it.

Like Wittenbrink et al. (1996), Andre and Cutler (1998) were also interested in developing functional applied uncertainty displays, but they focused on a much more complex environment aimed to mimic aviation and navigation. Unlike Wittenbrink et al.'s (1996) displays, their displays were dynamic and their experimental tasks were time sensitive (Andre & Cutler, 1998). Their first experiment aimed to assess differences in operators' risk bias and situational awareness both with and without uncertainty present. To do this, they created a sort of tracking task which had participants piloting a small virtual ship while both avoiding an asteroid while simultaneously staying as close to its proposed path as possible. In the no uncertainty condition, the meteors' path and position would appear unchanging, though in reality the actual meteor could be anywhere within an unseen area of uncertainty. In the uncertainty



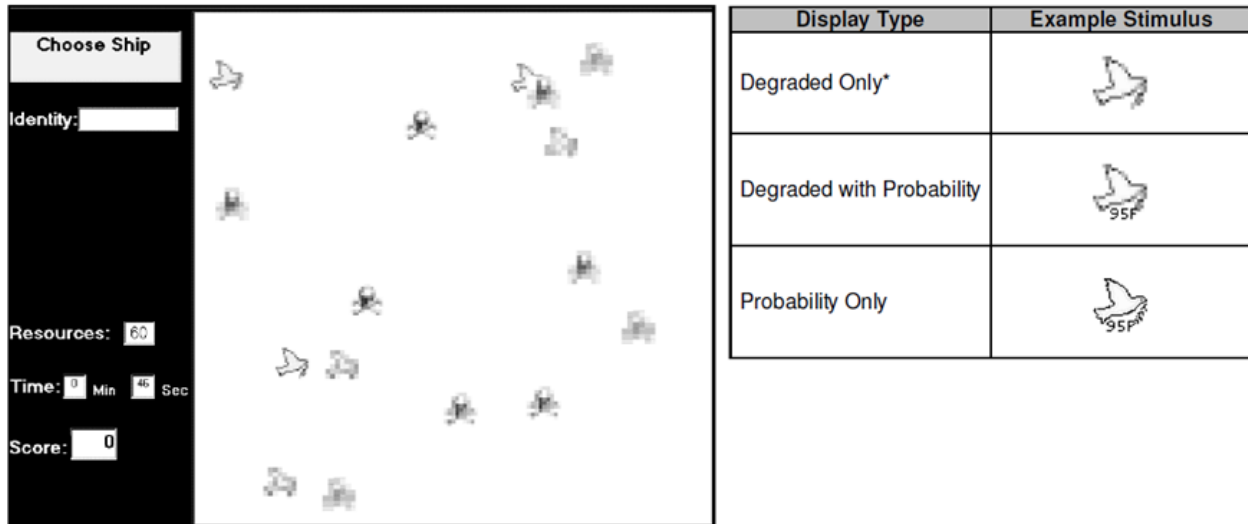
**Figure 5 - The left image is a still from Andre and Cutler’s (1998) second experiment depicting the gun turret at the bottom of the screen and the friendly and enemy ships flying above it. The right image depicts the uncertainty displays used in the study: the numeric, the graphical ring, and the graphical arc.**

display condition, uncertainty would be displayed to the user either numerically (i.e. you have a 10% chance of colliding with the meteor) or using a concentric circle with three different size/color combinations (see Figure 4). Results demonstrated that participants performed predictably poorly (measured by number of collisions) when the uncertainty inherent in the task was not displayed. Furthermore, despite being told the uncertainty was present (albeit not actively displayed), participants did not adopt a more conservative course of action (i.e. staying further away from the meteor and its’ associated path). This indicates that when uncertainty is not displayed, participants operate as if it is not present. When uncertainty was displayed, performance was significantly better than in the no uncertainty condition both when it was presented numerically and when it was presented graphically. However, performance in the visual display significantly outperformed the numeric representation.

In their second experiment, Andre and Cutler (1998) investigated the use of uncertainty representations to determine the tactical heading of aircraft. Their task consisted of putting

participants in control of a virtual gun turret with the aim of shooting down enemy planes while letting friendly planes pass unscathed. Uncertainty was represented in one of three ways, using a displayed number next to the planes, a graphical ‘ring’, or a graphical ‘arc’ (see Figure 5). Participants completed the task under three accuracy conditions: certain, moderately uncertain, and highly certain. In the certain and moderately uncertain conditions, participants shot down an equal number of friendly and enemy planes regardless of whether uncertainty was represented, or how it was represented. However, in the highly uncertain condition all uncertainty representations had better ratios (enemies killed versus friendly ships shot down) than the no uncertainty condition. Furthermore, the graphical arcs and rings performed better than the displayed number conditions. In addition, performance speed was faster in all uncertainty conditions, with the graphical arc slightly outperforming the other uncertainty representations, but not to a statistically significant degree.

When taken in their totality, Wittenbrink et al. (1996) and Andre and Cutler’s (1998) work provides strong evidence that uncertainty visualizations can be used effectively in complex environments, and the added information they provide can increase performance without a risk of degrading it in both static and dynamic environments. However, neither of these studies focus on what it is that makes a display inherently uncertain. Even when they do leverage existing mappings and representations to bolster their displays (i.e. Andre and Cutler’s (1998) use of red, yellow, and green to essentially represent uncertainty levels), they are not representations unique to the concept of uncertainty. To address this, Finger and Bisantz (2002) conducted a study investigating image blurring/degradation (a perceptually salient visual cue for uncertainty: see (Berlyne & Borsa, 1968; Joslyn & Rocha, 1998; Pang et al., 1997) for more on



**Figure 6 - The left image depicts the beginning of a trial in Finger and Bisantz’s (2002) decision making task. The right image depicts the three types of icons participants will be exposed to and how they are classified by the experimenters.**

this) to determine if leveraging these sorts of mappings could improve performance in a dynamic visual task.

Their study consisted of a decision-making task similar to Andre and Cutler’s (1998) second experiment. Participants were instructed to identify and classify objects as either friendly or hostile over a series of trials and were rewarded for correct identifications and penalized for incorrect ones. Objects were either doves or skulls and were either blurred/degraded, blurred/degraded with numeric probability, or not visually altered with numeric probability displayed. For doves, the various uncertainty displays (blurring, numeric probabilities, etc.) represented the likelihood an object was friendly; for skulls, the converse was true. For example, in the numeric probabilities conditions a ‘95’ next to a dove would represent a 95% chance that the object would be friendly, whereas the same number next to a skull would indicate the probability that it was hostile (see Figure 6). Participants’ goal was to correctly identify all



**Figure 7 - The representations used by Padilla et al. (2015) to represent uncertainty. The leftmost representation is a standard error bar, while the other representations were created for the purposes of their studies to evaluate their utility.**

objects (a total of 40 each trial) in a set time period without running out of resources. Correct identifications lead to a single resource being consumed (out of a total of 60) and points added to their total score, but incorrect identifications lead to twice as many resources being consumed and no scoring increase. Trials ended when time ran out, all resources were consumed, or all objects were identified (regardless of whether that identification was correct).

Results revealed significant learning effects in every condition, with performance increasing steadily across trials. Though accuracy was not significantly different across display types, performance for all other metrics (score, trial time, and number of identifications) was significantly greater in the blurring/degraded only condition. This suggests that leveraging psychologically salient and intuitive mappings can not only enhance performance, but does so at no cost to overall accuracy.

### **3.1 Lessons for Error Bars**

Taking past work into account, researchers attempted to integrate these lessons into the error bar representation (Padilla et al., 2015). Padilla and colleagues (2015) created a series of

visual glyphs/configural displays to determine whether non-experts could utilize them to make determinations of the overall accuracy of weather forecasts (See Figure 7). These glyphs were tested alongside the standard error bar representations to determine their overall accuracy and utility. Surprisingly, users tended to ignore these uncertainty representations altogether when making their decisions. They seemed to rely solely on the representations overall distance from the mean value displayed to make their decision. This spatial cue seemed to be the only one that users consistently deployed to make their decisions, ignoring all other aspects of the display. This is even more surprising in light of the fact that the authors were well versed in the research and leveraged color, size, blurring, etc. So, even if the information was being displayed accurately, clearly, and concisely; it was being ignored.

This is a problem that was inadvertently addressed in a study by Correll and Gleicher (2014). In an effort to discourage the use of error bars in favor of their own representations, Correll and Gleicher tested them against their own creations in a series of 3 experiments. They tested traditional error bars against their displays which varied in color, size, and blurring/degradation level (again, all previously researched psychologically salient measures). Participants were asked a series of questions across three experiments to determine whether they were able to effectively use the displays to answer them accurately. Experiment 1 asked them to make a neutral judgement (i.e. how likely is x outcome) and utilized only the traditional error bar vs textual descriptions of uncertainty, experiment 2 asked them to make a judgement of a political candidate's likelihood to receive x percentage of a vote (i.e. based on this display, how likely is candidate A to get 55% of the vote), and experiment 3 put candidates polling numbers

and uncertainty representations side by side and asked participants to make judgements about them (i.e. how likely is candidate A to win the election, etc.).

Results of experiment 1 demonstrated that the participants' judgements of uncertainty were far more accurate when error bars were present than when uncertainty was described in text (91.6% vs 62.2%). Though this result alone is not surprising (as it has been shown in countless past research), participants level of confidence in their own accuracy was surprising.

Participants were significantly more confident that they were correct when uncertainty was described textually than when it was shown on an error bar. This suggests some intriguing possibilities. However, first and foremost it suggests that participants are confident that they understand both bar graphs and text descriptions, but have much less confidence in their ability to use an error bar. It also demonstrates that confidence does not translate into accuracy.

Results of experiments 2 and 3 showed a mean difference, but no significant performance difference between error bars and the other display types when it came to accurately determining basic differences. However, when it came to inference of effect size (i.e. the magnitude of the perceived difference) the alternative gradient and violin plots outperformed all other displays. These plots integrate probability density functions (pdfs) into the displays either by varying the gradient of the display accordingly or by displaying the pdfs themselves rotated 90 degrees and symmetrically displayed around the mean. In addition, participants remained less confident in their accuracy for all uncertainty visualizations (regardless of type) than they did viewing a standard bar graph with text.

This is intriguing because it suggests that when participants are forced to use uncertainty representations to make accurate decisions that they can do so, but are not confident in their

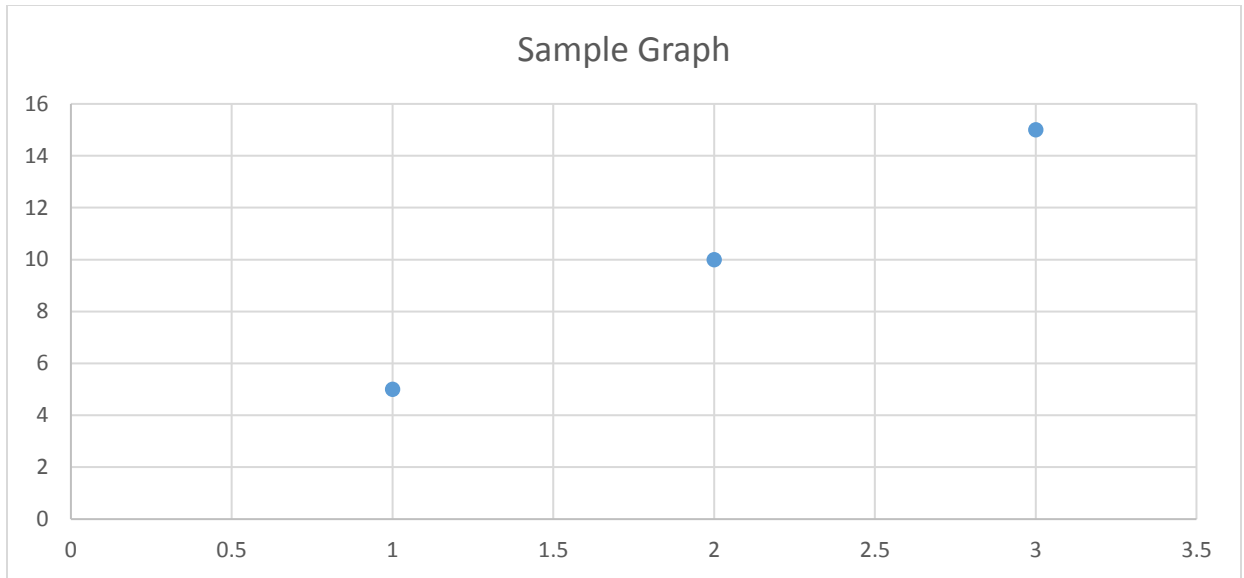
abilities. Furthermore, it is entirely likely that error bars not being taught or demonstrated could lead to an unfamiliarity that strongly effects participants confidence.

## **CHAPTER 4. AUDITORY REPRESENTATIONS OF MATHEMATICAL DATA**

Even if an ideal visual uncertainty representation could be created, understood, and attended to, this would still leave uncertainty representations completely inaccessible to a large portion of the population. As of 2015, there are over 60,000 students currently enrolled in American schools that meet or exceed the legal definition of blindness ("Distribution of Eligible Students Based on the Federal Quota Census of January 6, 2014 (Fiscal Year 2015)," 2016). These students rely on a combination of haptic and auditory displays (in addition to large print when applicable) to learn mathematics. Unfortunately, most of the research and development into these tools goes towards creating these displays for K12 students (Davison, 2013; Hansen, Liu, Rogat, Hakkinen, & Darrah, 2016). Because of this, more advanced mathematical displays remain entirely inaccessible. Though haptic displays are useful, they are expensive to create, often unmalleable, and difficult to transport. Because of this, sonification remains the most viable way to communicate mathematics displays to this population. As sonification is a vast field of research spanning decades, it is important that this discussion is constrained only to how to represent uncertainty in the context that it needs to exist in: the auditory graph.

### **4.1 'Standard' auditory graphs: the basic architecture**

Though there has been a tremendous amount of research on auditory graphs as a whole, there is still no real codified set of principles for how an auditory graph 'should' be structured (outside of the basic mapping of loudness to data values in one dimensional processes such as audification). This is hardly surprising though, after all, such a set of principles does not truly exist for visual graphs either. Like most things in science, how an auditory graph is structured



**Figure 8 - Sample visual graph to be sonified. Auditory equivalent explained in text with sound file available upon request.**

truly depends on what data it is representing and often (though this is not always the best strategy) which type of visual graph it is attempting to mimic. Because the stated focus of this paper is to create a framework for an uncertainty representation within the bounds of an auditory graph, what I am going to refer to as the ‘standard’ auditory graph is the sonification equivalent of a series of data points plotted on an x-y axis. This type of graph is not only prevalent throughout scientific publications, but also in statistics classrooms all over the world. For an auditory uncertainty representation to be functional, it needs to be able to exist within this structure.

#### **4.2 The rules of the game: what works and what does not**

The ‘rules’ for the standard auditory graph have come about as a result of a tremendous amount of work in a variety different environments, conditions, equipment levels, etc. (Flowers, 2005; Flowers, Buhman, & Turnage, 1997; Kramer et al., 2010; Walker & Mauney, 2010).

Flowers (2005) summarizes these ‘rules’ by conducting an overview of the research conducted on auditory graphs up until the articles’ publication and listing off a general dos and don’ts list. For what we know works, Flowers (2005) put pitch coding and time at the top of the list. These two principles are interlinked, and are the cornerstones of auditory graphs creation and interpretation to this day. Put plainly, it translates to using pitch to represent the values on the y axis and time to represent the x axis. For example, let’s say we have a scatterplot with three points, each one placed further along the x axis and higher on the y axis than each previous point (see Figure 9). The auditory graph generated from these points would separate them in time based on their relative distance from each other on the x axis, and in pitch based on their distance on the y axis. The resulting graph would sound like three ascending tones and research has shown that it and displays like it would be easily understood even by untrained participants (Bonebright, 2005; Bonebright & Miner, 2005; Bonebright, Nees, Connerley, & McCain, 2001; Kramer et al., 2010; Oren, Harding, & Bonebright, 2008; Peres & Lane, 2003).

In addition to pitch coding and time, Flowers (2005) also lists using relative loudness to accentuate critical events and points of interest and differentiating timbres (the unique acoustic properties of a sound i.e., what makes a violin sound different than a saxophone) to prevent unwanted perceptual grouping and facilitate grouping of like tones. These principles are very well supported by the literature and not disputed by practitioners even now, more than ten years after the articles initial publication (Nees & Walker, 2007, 2009; Walker, 2013; Walker & Mauney, 2010; Walker & Nees, 2011).

In terms of what does not work, Flowers (2005) put simultaneous plotting of continuous variables at the top of the list. Understanding and interpreting multiple auditory streams simultaneously has been something that has been argued and attempted for years with varying

degrees of success (Bregman, 1994; Hermann & Hunt, 2004; Schuett & Walker, 2013). For the most part, Flowers (2005) is correct that multiple stream interpretation is much more difficult to implement and understand than singular data streams (something that holds true in visual displays as well); however, recent research has shown that it may be much more feasible than previously thought (Schuett, Winton, Batterman, & Walker, 2014). Flowers (2005) continues on to state that if practitioners do choose to create displays with multiple continuous auditory streams, they should endeavor to use as few as possible and at least separate them by timbre.

Flowers (2005) also advises against using loudness changes to represent important continuous variables. This is extremely important, especially in light of the effectiveness and ubiquity of pitch coding. Because of how humans perceive sound in general, higher pitches are perceived by the listener as being louder and conversely lower pitches are perceived as softer. If loudness were used in concert with pitch coding in a continuous fashion, users would be unable to decouple the two successfully, leading to confusion and misunderstanding. This does not mean that loudness cannot be leveraged successfully in auditory graphs (as mentioned earlier, it is excellent for highlighting critical events), but it does mean it should be avoided as a primary perceptual cue for representing continuous variables, especially when there are superior options available (i.e. pitch and time) (Hermann, 2008; Kaper, Tipei, & Wiebel, 1999; Kramer et al., 2010; Madhyastha & Reed, 1995; Walker & Mauney, 2010; Walker & Nees, 2011).

#### **4.3. The conceptual model of auditory graphs**

On a basic psychophysical level, Flowers (2005) does an excellent job of summing up what should and should not be included in an auditory graph for it to serve as a functional sonification. However, he does not address how auditory graphs are processed on a more psychological and theoretical level. To rectify this, Nees and Walker (2007) organized the

literature on auditory graph perception into a conceptual model to give practitioners a more theoretical and data driven overview to guide the development of effective displays in the future. The model is split into four parts: task, listener characteristics, display characteristics, and environmental concerns. The authors stress that any successfully crafted auditory graph should take each of these components into account, and neglecting any of these aspects could lead to a failed display.

#### *4.3.1. The task component*

Nees and Walker (2007) operationally define the listeners' 'task' as the information the listener wishes to extract from the graph. Though this is a broad definition (as auditory graphing tasks can be anything from getting a broad overview of a dataset to estimating a single point), it is important to note that starting broad allows the model to cover more ground initially before delving deeper into specifics. The task component of the model is based primarily on Jones' Rhythmic Theory (Jones, 1976). This theory suggests that sequences of sounds (i.e. auditory graphs) could be described in terms of three types of relationships: nominal, ordinal, and interval.

##### *4.3.1.1. Nominal relationships*

Jones (1976) describes nominal relationships as those in which a listener can determine whether a set of sounds are the same or different. For example, nominal relationships can be seen as akin to a speech sound discrimination task. These tests present participants with a series of 'nonsense' syllables to determine whether they can notice any difference between the two (note: these assessments are often used to assess speech development in children) (Beving & Eblen, 1973). In terms of auditory graphs, this would be the extent to which a listener could identify whether the individual data points or trends they were listening to were identical or not.

Like the ‘nonsense’ syllables in the speech task, the direction, type, extent, etc. of the differences are not important, only that participants are able to determine whether or not they are the same.

#### *4.3.1.2. Ordinal relationships*

Ordinal relationships describe when listeners can determine not only whether sounds are the same or different, but also the direction of the difference. According to Jones (1976), an ordinal relationship is present when a sequence of sounds can be discerned by whether they are higher or lower pitched than one another. Though her theory specifies differences in pitch, it is reasonable to extend this to other auditory components that scale similarly. For example, auditory graphs leverage both pitch and time to present data optimally. Like pitch, audio presentation can vary in time and it makes sense to describe differences in presentation speed (i.e. whether sounds are presented faster or slower or another) as ordinal relationships as well. This should also extend to more nuanced auditory concepts such as attack, decay, and numerous others (Walker, 2002; Walker & Cothran, 2003; Walker & Kramer, 2004; Walker, Kramer, & Lane, 2000; Walker & Nees, 2005c, 2011).

#### *4.3.1.3. Interval relationships*

Interval relationships are those where listeners can not only determine the direction of a difference, but its magnitude as well (Jones’ again uses pitch in her examples, but this again can be extended to other auditory components). These are the most specific of Jones’ (1976) proposed relationships, but also the most useful for understanding more complex auditory tasks. These sorts of relationships are the relative gold standard for auditory display because they both offer the most specific information while also encompassing the other two relationship types (nominal and ordinal). However, despite the usefulness of these sorts of relationships, they have been the most difficult to achieve in auditory graphing tasks, especially point estimation (which

remains a heavily researched, but unsolved problem) (Metatla, Bryan-Kinns, Stockman, & Martin, 2015; Nees & Walker, 2008; Smith & Walker, 2005; Walker & Nees, 2005b).

#### *4.3.1.4. Theoretical predictions*

Because Jones' (1976) theory can be so easily applied to auditory graphs, Nees and Walker (2007) argue for it as the basis for their theory's task component. They note that if applied to auditory graphs, Jones' (1976) theory scales directly as expected. For example, as each of her three relationships builds on one another, she argues that this should make tasks involving the simpler relationships easier than those requiring the more difficult ones. This pattern is demonstrated clearly in the auditory graphs literature as discrimination between trends and data points is simple even for untrained participants (a nominal relationship), as is discrimination with direction (i.e. whether a trend is going up or down or one point is higher than another) (an ordinal relationship) (Flowers, 2005; Nees & Walker, 2007; Peres & Lane, 2003, 2005; Stockman, Nickerson, & Hind, 2005). However, as mentioned in the previous section, point estimation and other more complex interpretation and comparison tasks (interval relationships) are not as successful (Nees & Walker, 2008; Smith & Walker, 2005; Walker & Nees, 2005b).

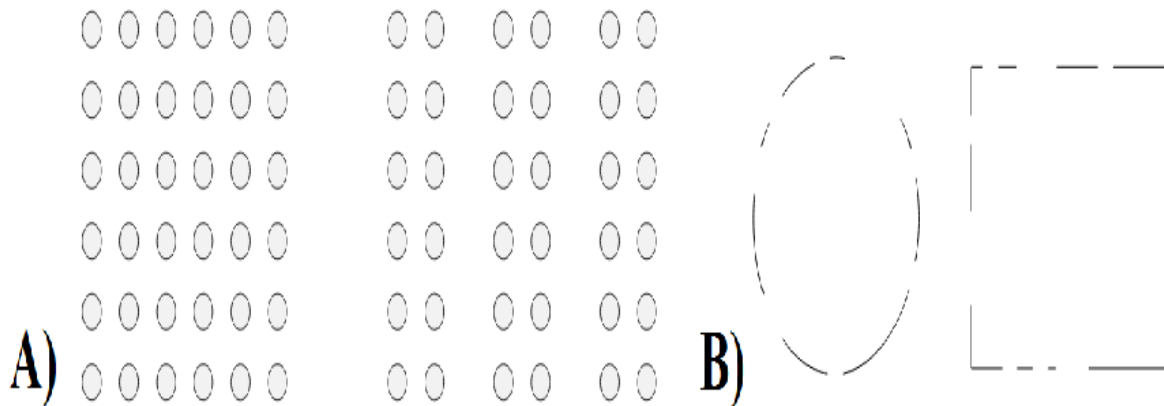
### **4.4. Listener characteristics**

Decades of research have provided a tremendous amount of knowledge about the human auditory system. With this in mind, Nees and Walker (2007) incorporate these various characteristics into their model, discussing the various commonalities and individual differences that can affect listeners' perceptions. They also delve briefly into training and learning and their potential effects on display perception.

#### *4.4.1. Commonalities*

Though people vary in a wide variety of ways, our basic biology is the same. In this sense, auditory perception at its simplest level is the same for everyone. Without delving too heavily into the biological underpinnings, what we perceive as sound is the transduction of vibrations in an audible range. Though the ranges can vary at their upper and lower limits, the most sensitive regions are fairly universal and often leveraged by good designers to optimize their displays. For example, people can generally differentiate between two pitches if they are between 1000-5000 Hz (barring tone deafness) regardless of age, training, musical ability, etc. (Kramer, 1993; Rossing & Stumpf, 1982; Walker et al., 2000). In addition to these psychophysical realities, Nees and Walker (2007) stress the importance of similarities in auditory grouping principles. These principles are derived from several fields of research (including Gestalt psychology, work on emergent features, default encoding, and others) with the singular goal of determining how sounds are grouped together into meaningful representations.

Though some auditory pattern perception theories are vague, the ones that are most useful are those that are able to make concrete and testable claims. Citing this, Nees and Walker (2007) return to discussing Jones' work, referring to her theoretical predictions as examples of the type of perception theory one should strive to achieve (Jones, 1976). Using the framework of her rhythmic theory, Jones (1976) made solid predictions that her described relationships (nominal, ordinal, and interval) would increase in perceptual difficulty, something that has held true in auditory graphs research (Flowers, 2005; Kramer et al., 2010; Peres & Lane, 2003; Stockman, Nickerson, et al., 2005). In addition to Jones (1976), other researchers have predicted that the



**Figure 9 - Image A is an example of the gestalt principle of proximity. The dots on the left portion of the image are very close to each other so they are perceived as a single object. The dots on the right portion are further apart so they are perceived in 3 separate groups. Image B shows two shapes with visible portions missing that through the principle of continuation are viewed as single contiguous objects.**

Gestalt principles of proximity and continuation would facilitate understanding of tone sequences (Deutsch & Feroe, 1981).

Originally proposed to explain the visual perception, the principles have since been extended to other senses including touch, and of most interest to this paper: audition (Chang, Nesbitt, & Wilkins, 2007; Cuddy & Lunney, 1995; Deutsch & Feroe, 1981; Wertheimer, 1923). Though several principles were proposed, the ones most relative to this paper are the aforementioned proximity and continuation. Visually, proximity refers to the fact that visual objects that appear closer together are presumed to be grouped together (see Figure 10a). Auditorily, this principle applies to both pitch and time. For example, two tones may be perceived as grouped together if they played successively and/or if they are close in perceived pitch. Conversely, separating tones along these dimensions will lead to them being viewed as

disparate elements (Cuddy & Lunney, 1995; Deutsch & Feroe, 1981; Kramer, 1993; Nees & Walker, 2007).

Continuation is one of the most important Gestalt principles, especially as it relates to design. Visually, continuation refers to the perception of overlapping or intersecting objects as uninterrupted figures (see Figure 10b). Auditorily, it refers to the continuity inferred in auditory patterns such as a systematic increase in pitch or tempo (Bregman, 1994; Kramer, 1993; Metatla, Bryan-Kinns, & Stockman, 2012). For example, a series of descending tones that are close in proximity will be perceived as a single descending scale (Walker & Nees, 2011).

#### *4.4.2. Individual differences*

Though a plethora of commonalities exist, auditory graphs can be perceived very differently by certain individuals (or groups of individuals) for a variety of reasons. The most basic reason for this is a difference in sensitivity. For example, as a general rule, younger adults and children have much better hearing than older adults, due simply to the normal deterioration the auditory apparatus undergoes with age. This is primarily due to cochlear decay, but may affect other structures as well (primarily in the inner ear, though auditory nerve damage can also occur) (Schnupp, Nelken, & King, 2011). Hearing damage/loss is also common in musicians, military veterans, airport tarmac workers, and others who are exposed to the extreme ends of the auditory spectrum (Kryter, 2013; Kryter, Ward, Miller, & Eldredge, 1966). Perceptually, these decrements lead to the inability to perceive sounds at the extremes of the auditory spectrum (i.e. if they are too soft/loud, high/low, etc.) as well as differentiate between tones that are too similar across the same dimension (Nees & Walker, 2009; Walker et al., 2000; Walker & Nees, 2011).

In addition to decrements, individuals can also excel at auditory processing. Frequently, musical ability is cited as a predictor of performance on auditory displays, but results have been

inconclusive. Some studies have suggested that musicians excel at understanding and interpreting auditory displays, citing slight performance increases when participants claim prior musical experience (Schuett & Walker, 2013; Walker & Nees, 2005b, 2011). However, these sorts of findings have been inconsistent at best, with significant differences being rare and difficult to replicate (Kidd, Watson, & Gygi, 2007; Nees & Walker, 2007; Walker & Nees, 2005c). Despite these inconsistencies, it is difficult to believe that an expert musician would not outperform a layman in interpreting audio. The reason for the minimal and inconsistent findings could be that there really is not an agreed-upon way to assess musical ability, especially within the bounds of a controlled research study. Because of this, researchers are forced to rely on the tenuous nature of self-reporting to draw their conclusions. Furthermore, participants are rarely asked to evaluate their own skill, but instead how long they have played an instrument, how many instruments they play, etc. (Batterman & Walker, 2012a; Nees & Walker, 2007; Schuett & Walker, 2013; Walker & Nees, 2011). It could be that self-assessment of skill may be a better way to gauge musical ability than years of experience, but evidence from self-assessment studies in other domains suggests that that too is fairly unlikely (Blanche & Merino, 1989; Davis et al., 2006; Yarbrough, 1987). It may be that if a strong link between musical ability and auditory graphs comprehension exists, discovering it may need to wait until a reliable measure is developed.

Individual differences in cognitive ability have also been investigated as potential links to auditory graph interpretation. Some evidence suggests that there are links between auditory graph comprehension, spatial ability, and verbal memory, with studies finding relationships between task performance along each of these variables (Kane et al., 2004; Nees & Walker, 2007; Trickett & Trafton, 2006). In addition, auditory graph comprehension has also been linked

to more generalized measures of fluid intelligence, notably Raven's progressive matrices (other measures, including the n-back test were also administered, but no significant links were found) (Walker & Mauney, 2004). Less validated intelligence measures (such as SAT scores) have also been investigated, but when significant links were found, the effects were very small, with test scores accounting for 5% or less of the variance in graph comprehension (Kidd et al., 2007).

Finally, evidence suggests that individuals can differ on sound mappings of conceptual data dimensions. These differences are sometimes based on severity and age of onset of visual impairment, but can also just be due to personal preference (Mauney, 2006; Walker, 2002; Walker et al., 2000; Walker & Lane, 2001). Put plainly this means that the auditory perceptual dimensions (i.e. pitch, tempo, etc.) that are used to represent a concept can mean different things to different people. For example, an increase in pitch could mean an increase in temperature to some, but a decrease to others. These concepts will be explored in depth in section 3.4.2, but because it is a way individuals can differ in auditory graphs comprehension, it was necessary to mention it briefly here as well.

#### *4.4.3. Training and learning*

Of all of the listener characteristics discussed thus far, training and learning has the smallest overall body of research to draw from. This may seem strange to many psychologists because as a whole training and learning has an extremely rich and varied literature, but this literature does not extend to auditory graphs comprehension (Nees & Walker, 2007; Walker & Nees, 2005b, 2011). There are several reasons for this (including the relative novelty of the display to most participants, how data in these studies is collected, etc.), but what research has been conducted has been extremely promising.

In point estimation tasks, Walker and Nees (2005) investigated which types of practice, if any, yielded the least amount of participant error. Participants were assigned to receive either a non-relevant filler task, practice only (which consisted of completing the task a set number of times without feedback), practice with feedback, practice with a visual prompt (where the equivalent visual graph was displayed to participants as they practiced), and conceptual training (a detailed instructional program explained in depth in Smith and Walker (2005), but that consisted of breaking down the task step by step and explaining each portion). Results demonstrated practice with feedback was superior to all other conditions and was only condition that was statistically different from the filler task. This illustrates a pattern that has been found in previous perceptual learning research showing that feedback presented in the same sensory domain for the same task will increase performance (Annett, 1966; Annett & Paterson, 1967; Walker & Nees, 2005a, 2005b, 2005c). Though these results are encouraging, they only scratch the surface of the potential training has to effect auditory graphs comprehension. Absent training, participants in auditory graphs studies already perform at high levels, so with it, it is possible that the performance ceiling is much higher than has already been measured (Flowers, 2005; Kramer et al., 2010; Nees & Walker, 2007; Walker & Nees, 2011).

#### **4.5. Display characteristics**

The next variable grouping in Nees and Walker's (2007) model are the characteristics necessary to actually craft the bottom-up elements of an auditory graphs display so that it can be understood and interpreted in its intended fashion. In contrast with section 3.3.3 (training and learning), this section has the most research to draw from, citing notable work from psychology computing, music technology, and more (de Campo, Frauenberger, Vogt, Wallisch, & Dayé,

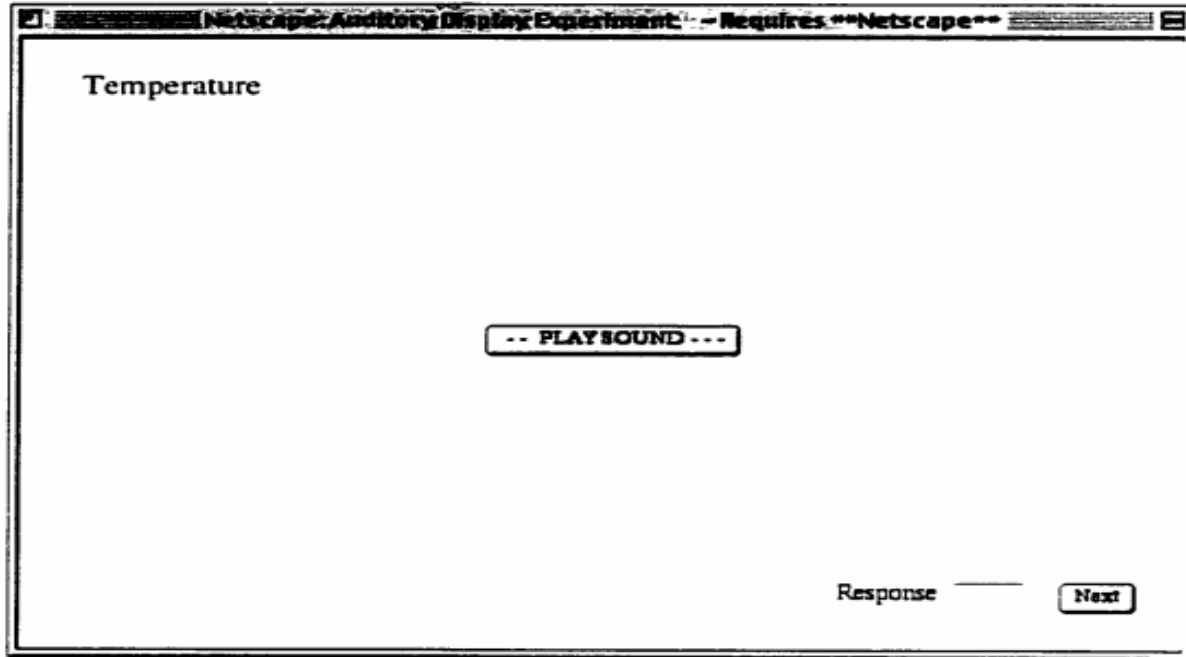
2006; Flowers, Whitwer, Grafel, & Kotan, 2001; Metatla et al., 2015; Väljamäe et al., 2013; Walker & Nees, 2011).

#### *4.5.1 Data*

Data are necessary for any quantitative display, and auditory graphs are no different. Though data comes in a variety of forms, the most important distinction that needs to be made is whether the data set is static or dynamic. The majority of data sets are static, where finite values are recorded and stored in a database or similar structure. However, dynamic sets are becoming more and more common as certain types of data are presented as they are collected in rapid succession. For example, a user could be looking at changes in stock pricing over the last quarter (a static data set) or actively following the fluctuations of the market during a trading day (a dynamic set). Furthermore, the nature of the data itself (i.e. what it is actually representing on a conceptual level) is also of the utmost importance and will be explored in detail in the following section.

#### *4.5.2. Mappings, scalings, and polarities*

Of all the components discussed thus far, these three may be the most important. Failure to execute any of them correctly will not only lead to a failed display, but one that represents the opposite of what it is intended to. Mapping refers to which auditory dimension covaries in tandem with which component of the data represented by the auditory graph (Nees & Walker, 2007). Changes in data can be mapped to anything from changes in pitch, to changes in tempo, loudness, brightness, etc. Furthermore, dual coding along these mapped dimensions can lead to performance increases due to redundancy gain (Walker, 2007; Walker & Lane, 2001; Wickens, Hollands, Banbury, & Parasuraman, 2015). Despite the dimensional variation, research has



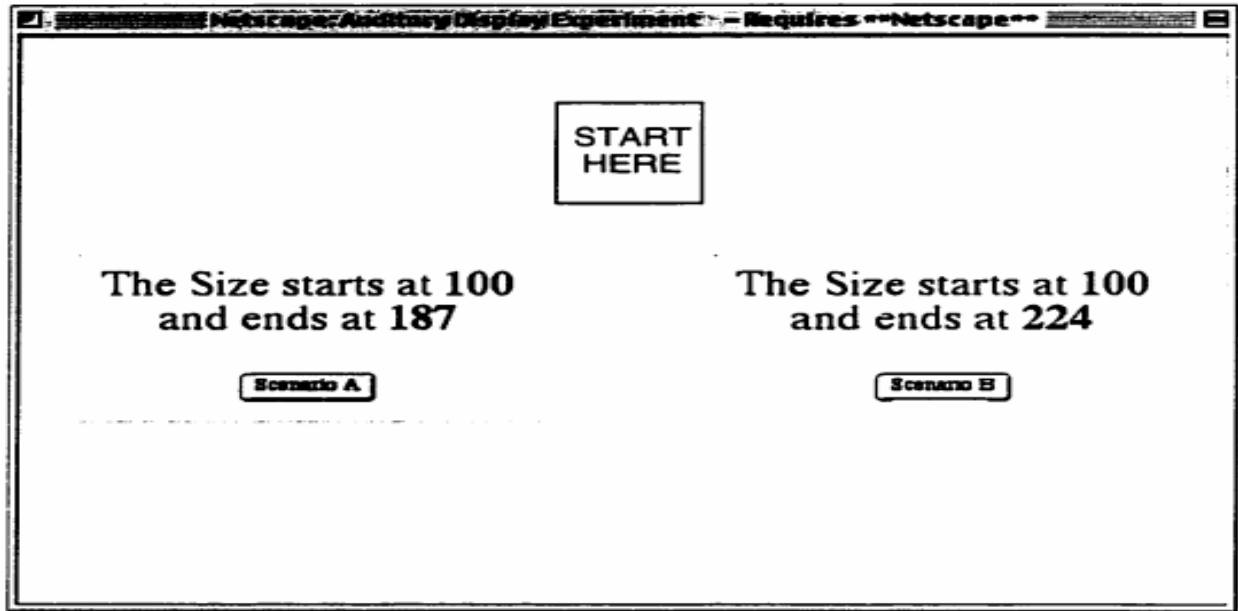
**Figure 10 - An image of a trial from Walker's (2002) conceptual magnitude estimation study. The conceptual dimension (in this case temperature) is on the top left, the button in the center plays the audio, the response box in the bottom right is left blank for the participant to assign a number to the auditory stimulus, and the next button moves them on to the next trial.**

focused primarily on frequency (perceptually pitch) mappings for a variety of reasons. Not only does it fit well into existing theoretical models (e.g. Jones' rhythmic theory), but it is also the one that allows auditory graph creators the most amount of leeway (Nees & Walker, 2007).

However, despite this, auditory graph developers usually stick to the MIDI piano timbre over the other tools afforded to them. Research has demonstrated that these frequencies are pleasant, within the audible range for the majority of people, and preferred to and outperform graphs that utilize pure tones (S. A. Brewster, Wright, & Edwards, 1993; Brown, Brewster, Ramloll, Burton, & Riedel, 2003; Nees & Walker, 2007, 2009; Walker & Nees, 2011). Though MIDI has been

the standard for a while, higher fidelity auditory equipment (both hardware and software) could lead to it being supplanted by more realistic synthesizers in the future.

Scaling is the extent to which a unit of change in a conceptual data dimension relates to a certain amount of change in an auditory dimension (Nees & Walker, 2007). Scaling in auditory graphs research is primarily based on the conceptual magnitude estimation procedure developed by Walker (2002). Walker's (2002) procedure was based on the traditional magnitude estimation procedures prevalent in psychology for years, but with a more cognitive and conceptual approach (Logue, 1976; Stevens, 1975; M. Teghtsoonian & Teghtsoonian, 1971, 1983; R. Teghtsoonian, 1971, 1973). Taking cues from Teghtsoonian and Tehtsoonian (1983) who used magnitude estimation procedures to successfully and repeatedly relate values to lines and shapes, Walker (2002) aimed to do the same with concepts and sounds. To do this, he presented participants with auditory cues varying on a single dimension and asked them to relate that to levels of a given concept. For example, participants would be told that the sounds they are going to be presented with represent temperature and to assign them numbers accordingly (see Figure 11). The first sound they hear would be given a pseudorandom number that participant feels the concept represents, but every subsequent sound would be directly related to it. So, if the participant thought the first sound represented a temperature of 10, they would directly scale the next sound they heard to that one. If they thought the stimulus sounded half as hot, they would assign it a 5, twice as hot a 20, etc. These values were then used to create a series of scaling equations unique to each auditory dimension and each concept. Returning to temperature, this means that there would be a scaling equation relating changes in temperature to changes in frequency, and another relating it to changes in tempo. Because frequency is directly linked to loudness (i.e. individuals perceive higher pitched sounds as louder and lower pitched sounds as



**Figure 11 - An image of a trial from Walker's (2002) follow up study assessing the efficacy of conceptual magnitude estimation in a more applied context. Clicking the 'START HERE' button will play the series of sounds and then participants choose which numeric range they feel best represents the concept of size by clicking the 'Scenario A' or 'Scenario B' buttons.**

softer), all stimuli were corrected for equal loudness using phon curves (Robinson & Dadson, 1956; Schneider, Wright, Edelhert, Hock, & Humphrey, 1972; Suzuki & Takeshima, 2004). A phon is a unit of perceived loudness and correcting stimuli to correspond to phon curves entails making higher pitched stimuli softer and lower pitched stimuli louder to the point where they are perceived as perceptually equivalent (Robinson & Dadson, 1956; Schneider et al., 1972; Walker & Lane, 2001). Correcting stimuli in this manner ensures that any scaling equation developed for a given perceptual dimension would be based purely on changes in frequency and not biased in any way due to perceived loudness.

Once scaling equations were obtained for a variety of conceptual dimensions (ranging from temperature, to size, to pressure, etc.), Walker (2002) derived values from them in order to test whether they held true in a more applied context. This was done using the following equation (or a variant):  $P_5 = P_1 \times (F_5/F_1)^m$ . P in this case represents the concept of pressure, with  $P_1$  representing the starting pressure for the lowest frequency stimulus and  $P_5$  representing the final pressure.  $F_1$  represents the frequency of the starting stimulus and  $F_5$  the frequency of the final stimulus, with m representing the slope of the scaling equation. Stimuli were then created by standardizing a starting point (in this case 100), calculating what the end point should be using the scaling equation, and then pairing that range of values with values that were 20% higher or lower than it (see Figure 11 for an example of this using the conceptual dimension of size). These 20% altered values were also paired with each other in a series of ‘catch’ or ‘foil’ trials to detect any potential biases participants may have towards higher values, lower values, clicking the left value over the right, etc. The auditory stimulus in a given trial would consist of a series of five tones started at a set base value ( $F_1$  in the example equation) and ending at the value predicted by the scaling equations for that range ( $F_5$  in the example equation). Participants would then select which range they felt the auditory stimulus set best represented. Results from the study indicated that participants selected the ranges predicted the various scaling equations significantly more than would be predicted by chance for each conceptual dimension tested and for both variations in frequency and tempo.

In addition to mappings and scalings, polarities are also of the utmost importance. In auditory displays, polarities represent the relationship between the conceptual mappings and the direction of the auditory dimension being manipulated. So, for example, for temperature, participants almost universally feel that higher pitches represented higher temperatures (Walker,

2002; Walker et al., 2000; Walker & Lane, 2001; Walker & Nees, 2005c). This sort of relationship is referred to as a positive polarity. By contrast, research has shown that size has a negative polarity with lower pitches felt to represent larger sizes (Smith & Walker, 2002; Walker, 2002). Mapping a concept with an incorrect polarity can not only lead to a display being misunderstood, but in many cases understood in the exact opposite of its intended meaning. Furthermore, polarities can differ amongst different aspects of the population, making identifying accurate polarities an even more crucial aspect of not only auditory graphs design, but auditory display design in general. For example, when assessing polarities for money (i.e. what change in which auditory dimension represents more money), Walker and Lane (2001) found that participants who were congenitally blind (i.e. blind from birth) had opposite mappings and polarities than individuals who were sighted or became blind later in life. For the non-congenitally blind participants higher frequencies represented more money, but for the congenitally blind the opposite was true. These participants were asked after the initial portion of the study what their reasoning for this was, and they gave an extremely interesting answer. They said that if you took a single coin and dropped it on a table, it would make a high pitched ‘plink’ noise, but if you dropped a large sack of coins it would make a low pitched ‘thud’. Though these sorts of systematic mapping differences are rare, they are important to keep in mind not only when designing auditory graphs, but for auditory displays in general.

#### *4.5.3. Context*

For auditory graphs, context refers to everything in the display besides the rote representation of the data that is used to facilitate comprehension (Nees & Walker, 2007). Early visual graph comprehension theory had a limited focus on these concepts, but later variants of it have shifted their focus to recognize the importance of these contextual elements (Cleveland &

McGill, 1984; Kosslyn, 1989; Pinker, 1990; Shah & Hoeffner, 2002). Auditory graphs research has followed a similar pattern, with early work ignoring these concepts and later work putting more focus on them (Frauenberger, Stockman, & Bourguet, 2007; Nees & Walker, 2006; Smith & Walker, 2002, 2005). Context has been shown to be most useful in more difficult auditory graphs tasks like point estimation. For example, Smith and Walker (2002, 2005) demonstrated that reference tones on the y axis and rhythmic clicks on the x axis can both facilitate the process. In addition, Nees and Walker (2006) showed that intensity changes can also help with other difficult auditory graphing tasks such as stream segregation, which would make differentiation between different data streams easier. Stream segregation is also facilitated by using different timbres for different types of data. For example, the main data stream the participant is supposed to attend to may utilize the traditional MIDI piano, but differing related streams may utilize synthesized versions of other instruments, or even non-musical sounds to convey their information and attempt to ensure that it is processed differently (Bregman, 1994). Though context is an important part of auditory graph design (especially as data sets get larger and more complex), there is still much more research needed to determine the extent to which context can positive influence auditory graphs comprehension (Nees & Walker, 2007).

#### *4.5.4. Temporal characteristics*

Put simply, the temporal characteristics of an auditory graph refer to the overall duration of the graph and the lengths of time taken up by its individual components (i.e. representations of single data points). At present, there is no ideal duration for an auditory graph or its' individual stimuli, but the general consensus is that graphs should not exceed 10 seconds if at all possible (Flowers & Grafel, 2002; Flowers et al., 2001; Walker & Mauney, 2010; Walker & Nees, 2011). However, research has shown that the presentation rate of individual stimuli can have an effect

on performance depending on what information the user needs to extract from the graph and what they need to do with it. For example, evidence suggests that participants perform better in point estimation tasks when 1 or 4 data points were presented per second as opposed to another number (presentation rate ranged from 1 to 8 per second) (Nees & Walker, 2007, 2008). In addition, if presentation rate is too quick (i.e. 100 milliseconds or less) performance deteriorates if adjustments are not made to frequency and intensity to compensate (with adjustments, successful perception of stimuli as short as 25 milliseconds has been observed) (S. Brewster, Wright, & Edwards, 1995; Doughty & Garner, 1948; Goldstein, Öquist, & Björk, 2002; Nees & Walker, 2007; Turnbull, 1944; Walker & Nees, 2011). Furthermore, increased presentation rate leads to a greater number of tones being presented in a single sequence which can lead to a major perceptual drop off. Current research places the perceptual ceiling for auditory graphing tasks from 9 to 11 stimuli, but it is truly task dependent (Bianchi, Oakley, & Kwon, 2012; Cheatham & White, 1954; Garner, 1951; Nees & Walker, 2007; ten Hoopen & Vos, 1979; Walker & Nees, 2011).

#### *4.5.5. Multiple data series*

There has been a tremendous amount of work done on auditory presentation of multiple data series, but not much that specifically deals with auditory graphs. What has been done has primarily focused on using spatial separation and differing timbres to facilitate overall comprehension (Bregman, 1994; Brown & Brewster, 2003; Brown et al., 2003; Walker & Nees, 2011). In addition, research has focused on parallel versus serial presentation of auditory data streams, finding that which is preferable is mostly task dependent (Brown et al., 2003; Schuett & Walker, 2013; Schuett et al., 2014). Flowers (2005) has noted that for some tasks, especially graphing and mathematics, sequential presentation may be preferred over serial. His argument is

rooted in the observation that many displays using parallel audio presentation in this realm have been woefully ineffective. Furthermore, there have been several successful sequential displays in this space conveying information such as box plots, statistical graphs, and even exploration of complex data tables (Flowers & Hauer, 1992, 1993, 1995; Kildal & Brewster, 2005; Kramer et al., 2010; Peres & Lane, 2003; Stockman, Hind, & Frauenberger, 2005; Turnage, Bonebright, Buhman, & Flowers, 1996).

#### **4.6. Interactions and mutual influences of task, listener, and display characteristics**

Nees and Walker's (2007) conceptual model closes out with a discussion of the importance of how their various model dimensions (task, listener characteristics, and display characteristics) can interact with one another. In short, not much is known about the interaction between task, listener characteristics, and display characteristics and research examining such interactions is truly in its infancy. Some work has attempted to make this interaction easier to study by developing in depth sonification classification systems, but these have yet to be tested to the point that they can be deemed accurate and feasibly deployed (Grond & Hermann, 2014; Ludovico & Presti, 2016).

#### **4.7. Environmental concerns**

Though most experiments start in the lab, the eventual goal is to have them deployed in the real world in ecologically valid scenarios. Some have argued that the nature of psychophysical work and the tremendous amount of control experimenters have over stimuli, the situation, etc. endangers their potential real world application (Neuhoff, 2004; Watson & Kidd, 1994). However, this has been shown to not be the case on several occasions with auditory graphs and displays, with lab based psychophysical work being used to successfully develop and deploy educational graphing software in a classroom, a fantasy sports website for visually

impaired users, an accessible weather application, and more (Batterman, Schuett, & Walker, 2013; Y. C. Chew, 2014; Y. C. D. Chew & Walker, 2013; Davison, 2013; Flowers & Grafel, 2002; Flowers et al., 2001; Schuett et al., 2014). That said, environmental concerns play a key role in how researchers and designers go about creating effective displays and ignoring them could lead to a display that fails to function in its intended environment. For example, if a certain frequency is extremely prevalent in a given environment, it would be foolish to design a display that only emitted that frequency as it would likely get masked and fail at communicating its intended meaning.

#### **4.8. A brief commentary on split attention and the redundancy paradox**

Research in educational psychology and instructional design has largely supported the idea that presenting information in different modalities is beneficial to learners. However, there is some limited research that has found that redundant displays can split attention and decrease performance (I refer to this as a ‘redundancy paradox’) (Kalyuga, Chandler, & Sweller, 1999; Leslie, Low, Jin, & Sweller, 2012; Sweller, van Merriënboer, & Paas, 1998). Though these findings are interesting, they do not comport with the majority of work in both sonification and human factors that demonstrates the benefits of crossmodal displays (Bianchi et al., 2012; Bonebright, 2012; Kaper et al., 1999; Wickens, 1992, 2002; Wickens et al., 2015). In consideration of the preponderance of the research evidence, it is my view that well-designed multimodal displays can generally be expected to yield a performance gain, and as such multiple perceptual modes should be considered when designing optimal displays.

## CHAPTER 5. AUDITORY UNCERTAINTY

### 5.1 Introduction

Now that I have covered the history of uncertainty representations, the auditory graphs structure, and the conceptual model of auditory graphs, it is important to delve into work specifically on auditory uncertainty. Though very little work has been done on auditory representations of uncertainty, the most recent work on the subject has offered some new insights that are important to take into account and are essential to the creation of an ideal auditory uncertainty representation, especially one that fits into the existing auditory graphs structure (Batterman & Walker, 2012a, 2013).

### 5.2. Building from the ground up

Until recently, when auditory uncertainty representations were deployed (which was exceedingly rare) they were represented using forced and fairly arbitrary mappings (Ballas, 1994; Fisher, 1994; Lodha, Wilson, & Sheehan, 1996). For example, LISTEN was a system developed to communicate auditory uncertainty, but in actuality it was a basic sonification interface that offered users some control over how a dataset was sonified. Not only was there nothing truly ‘uncertain’ about the sonifications produced by LISTEN, but its functionalities were dwarfed by later more powerful sonification creation software like the Sonification Sandbox (Davison & Walker, 2007; Lodha et al., 1996; Walker & Cothran, 2003). This lack of research lead Batterman and Walker (2012) to essentially start from the beginning to try to determine what aspects of a sound truly make it ‘uncertain’. To do this, a wide array of sounds were piloted ranging from chords, to bent notes, to natural sounds, and even banging on a wide array of objects. However, what ended up working best was noise.

In acoustics, white noise refers to a sound that contains every frequency in the audible range of human hearing (20-20,000 Hz) at equivalent spectral density (Kryter, 2013; Mancini & Carter, 2009). Perceptually, it has been described as sounding like a waterfall, wind blowing, or since the advent of modern technology: static. It is this ‘static’ sound that pilot participants latched onto, citing its common link to perceptually uncertain events. Participants cited the presence of noise when they were stuck between radio stations, or channel surfing and encountering a weak or broken signal as examples. In order to determine whether noise could be viable as a representation of uncertainty, Batterman and Walker (2012) modified and replicated Walker’s (2002) conceptual magnitude estimation task using noise in lieu of the melodic tones present in the original task. In order to vary the noise so that it would have varying and discernable pitches to it, bandpass filters were created around the central frequencies used in Walker’s (2002) initial study. So, for example, instead of a playing a melodic tone at 400 Hz, participants would hear noise with a filter built around the 400 Hz frequency. This filter would systematically soften the frequencies higher and lower than 400 Hz so that the resulting sound would be akin to pitched static.

Replicating Walker’s (2002) task provided a series of scaling equations for both changes in frequency and changes in tempo, revealing that participants viewed increases in both as mapping to increases in uncertainty levels. Furthermore, the equations created had exceptionally high  $R^2$  values (over 0.95 in almost every condition) matching or exceeding the values Walker (2002) found in his work. However, though it had several successful replications, no one had ever attempted to use Walker’s (2002) procedure to map non-melodic tones, so even though the resulting scaling equations were promising the authors felt more validation was needed before

they could be confidently deployed as an uncertainty representation (Batterman & Walker, 2012a, 2013; Walker, 2007; Walker & Mauney, 2004, 2010; Walker & Nees, 2005c, 2011).

In order to validate the scaling equations, Batterman and Walker (2013) used the equations to generate value ranges similar to Walker's (2002) follow up to his initial conceptual magnitude estimation study and tested them in a similar fashion. Results demonstrated that participants chose the ranges produced by the equations at a significantly higher rate than would be predicted by chance, again replicating previous work with melodic tones (Batterman & Walker, 2013; Walker, 2007; Walker & Mauney, 2010; Walker & Nees, 2005c, 2011). This work demonstrated that not only is noise a valid and intuitive mapping for uncertainty, but that it performs equal to (and at times better than) universally accepted auditory mappings like those for temperature, size, and perceived urgency (Burt, Bartolome, Burdette, & Comstock Jr, 1995; Guillaume, Pellieux, Chastres, & Drake, 2003; Hellier & Edworthy, 1989; McNeer, Bohórquez, Özdamar, Varon, & Barach, 2007; Walker, 2007; Walker & Kramer, 2004). These findings become even more impressive when taking a deeper look at the stimulus creation in Batterman and Walker's (2012, 2013) work.

When creating the bandpass filters for their studies, Batterman and Walker (2012, 2013) used a 6 dB decay instead of 12 dB or even 18 dB. This means that around the central frequency, each octave was systematically softened at a rate of 6 dB. These types of filters are referred to as 'first order' filters and are fairly common in music, acoustics, and psychophysical research and are often used to modulate and control tones in commercial audio equipment (Ballou, 2013; "Equalisers Explained," 2001). However, first order filters have their limitations. In addition to the fact that the central (or cutoff if it's a high/low pass and not a bandpass filter) frequency will be less clear due to its' surrounding frequencies being louder, it is also unable to

resonate at a particular frequency. Normally this lack of resonance would not be too big of a deal, in fact commercial speakers are often calibrated with first order filters to purposefully reduce resonance (Ballou, 2013). However, next to noise (though noise was the overwhelming favorite), overly fluctuating or resonating tones were seen as the next best option (Batterman & Walker, 2012b). Because of this, adding resonance could further increase the efficacy of noise based uncertainty displays by leveraging the principle of redundancy gain (Wickens et al., 2015). Though this principle is primarily applied to visual displays, it has been shown to function to increase performance in auditory displays as well (Peres, Best, Brock, & Stockman, 2008; Peres & Lane, 2005). If future studies used second order filters in stimulus creation instead of first order ones, the higher Q rating would allow the central frequency to resonate and the 12 dB decay could make it easier to perceive and compare to other stimuli (Aldrich, Hellier, & Edworthy, 2009; Ballou, 2013; "Equalisers Explained," 2001). Doing this could lead to improvements over an already successful representation, leading to even greater understanding and bolstering its potential usefulness in future applications.

## CHAPTER 6. RESEARCH QUESTIONS

Taking the totality of this paper into account, one is left with a large number of unanswered questions:

R1: Can confidence intervals be better understood with alternative visual display types?

R2: Do questions involving confidence intervals have a unique status when it comes to the extent they are better understood with graphs present?

R3: Can the presence of these graphs enhance performance even further if uncertainty is also represented on them?

R4: Can confidence and performance be increased with prior training and explanation?

R5: Can use of other modalities [i.e. auditory displays] enhance student performance in addition to making the displays accessible?

To address this, the following studies were conducted.

## CHAPTER 7. METHODS

### 7.1 Study 1

The literature described to this point suggests that confidence intervals and uncertainty representations as a whole can enhance performance in statistical tasks, and that training may also play a role. The following study was conducted to determine this by presenting participants with questions about confidence intervals and probability (another difficult subject area that students often struggle with) along with varying display types and training information to see how this would affect overall performance. Participants' overall confidence was also assessed to determine both the effect of training and on performance.

#### *7.1.1 Participants*

Participants for this study were 37 Georgia Tech students and/or members of the surrounding community. They were recruited through the university SONA system or through word of mouth. Participants recruited through SONA received 1 credit per hour spent completing the study. They had normal or corrected-to-normal hearing and vision and completed either a first semester statistics course at an accredited university and/or the statistics AP exam.

#### *7.1.2. Experiment structure*

The purpose of this study was to address the first four proposed research questions. To do this, participants were asked to complete an abbreviated version (6 short answer questions rather than 5 short answer questions and a longer investigative task) of the free response section of the AP Statistics exam. Four of the questions were about confidence intervals and two were about probability, another topic frequently covered on the exam.

Probability was chosen because like confidence intervals there are many questions that

focus on it, it is also a topic that students have difficulty with on the exam (albeit not as much trouble as confidence intervals), and there is a large body of research on how it is understood and processed. For the confidence interval questions, one had no graph, one had a graph present, one had a graph with confidence intervals present, and one had a graph with an alternative uncertainty display drawn from Correll and Gleicher's (2013, 2014) work. For the probability questions, a graph was either present or absent. Each participant received one of each question type and be given an hour to complete them.

Prior to completing the exam questions, participants filled out a form detailing their level of confidence in their mathematical, statistical, and test taking abilities and half of them were given brief training on confidence intervals, their visual representations (both error bars and the alternative display), and how to interpret them. The other half of participants were given brief training on probability. Following this training/explanation, participants were again asked to fill out confidence measures.

### *7.1.3. Question Selection*

Questions were selected based on their viability within the context of this study. For a question to be viable, it must have appeared on a past AP Statistics exam and the presence of an uncertainty display cannot reveal the answer to the question. Furthermore, it must be possible to complete the question with or without the display present. This is to ensure that the displays are evaluated based on their usefulness and not their necessity. All questions and other experimental materials are included in the appendices.

### *7.1.4. Alternative Display Creation*

The gradient uncertainty display was created from Correll and Gleicher's (2013, 2014) work. Their display was used at the baseline and then the length, width, and degradation were

modified to fit the uncertainty levels described in the chosen test questions. The length was always the same as the error bar display and was superimposed over it in questions where error bars already existed, and created to match where the error bar would be in conditions where it was not present. The width was adjusted in accordance with the length to create a larger and more salient uncertainty representation. The magnitude of the display width began at the width of an error bar's horizontal brackets, and expanded or contracted according to the amount of uncertainty being displayed, up to a factor of 2 (meaning that the largest amount horizontal variation within the questions used for this study was twice the width of an error bar's horizontal brackets and the smallest was one half the width). For example, to create the uncertainty representation for a given set of data points where the error surrounding a given set of means was 0.5, 1, and 1.5, the width of the gradient display would be set to half the standard width for the first value, the standard width for the second value, and 50% larger than the standard width for the third value. This difference was larger in the training than it was in any of the chosen test questions to ensure that participants realized that the width corresponded with uncertainty just as much as the length did, as it was a display concept that participants were likely not familiar with.

The degradation levels varied similarly to the width, in that the display appeared more degraded the more uncertainty was being represented. 'Degraded' in this sense refers to the clarity (or lack thereof) in the gradient display. Clarity was altered by using a Gaussian blur filter on GIMP, with the maximum blur being set to size x and y values of 2.5. These values indicate the standard deviations for the horizontal axis and spatial scale factors respectively. They were chosen because they allowed for the maximum amount of blur while still allowing for the image to be clearly visible and interpretable. These blur filters were only placed over the 'tails' of the uncertainty display so that the dark center portion remained equally clear in every

instance. This was meant to communicate a sense of certainty that would be lessened by variable rates the further away it was from the center point.

#### *7.1.5. Procedure*

Prior to their arrival to the laboratory, participants filled out their basic demographics information online, in addition to affirming that they met the criterion to participate in the study. They were then randomly assigned to either receive confidence interval training or probability training. Upon arrival, participants received a consent form; and an approved calculator if they did not have their own. At this point they were given the confidence measures to fill out briefly before receiving their training packet. At the end of the packet, they were asked to fill out these confidence measures. Following this they were given the exam questions and materials (the same z, t, and chi square tables and equation sheets provided to students taking the AP exam) and told they will have one hour to complete the questions. When they completed them/when the hour ends, they were debriefed and thanked before leaving.

#### *7.1.6. Hypotheses/Expected Results*

This study evaluated the effect that different display types have on confidence interval questions and whether or not this is mitigated at all by training and exposure to them. It also assessed the confidence participants have in their own abilities, whether training can increase this, and if this increase correlates with superior performance. Though the study primarily focuses on confidence interval questions, two questions were on another topic (probability). This was to assess the impact visual graphs have on other statistical topics and compare them to confidence interval questions. It also allowed for comparisons between performance changes due to training.

Based on prior work, it was expected that confidence interval questions would perform

worse than questions on other topics. However, this difference was expected to be mitigated by both training and display type. Absent relevant training, scores should be highest on the other topic questions with visual display, followed by the other topic without display, followed by confidence interval with graph, followed by confidence interval with alternative display, and finally confidence interval with error bars. With relevant training, the result pattern should be confidence interval with alternative display, confidence interval with error bars, confidence interval with graph, other topic with display, other topic without display, and finally confidence interval with no display. Without relevant training, the observed pattern of results is expected to closely resemble the data analyzed in the AP Statistics exams of the past. Other topics outperform confidence intervals as a whole, and questions with readily understood visual displays score higher than those without them. In addition, displays that are not well understood (like error bars and alternative uncertainty displays) are ignored and their scores line up with those with no visual displays present. Participant confidence is expected to increase in the topics of the training they receive and that increased confidence is expected to correlate with increased performance. However, this increase should be greater for confidence interval questions as the baseline confidence level is expected to be lower.

## **7.2. Study 2**

Following up on Study 1, Study 2 leveraged existing literature to create and evaluate auditory error bars to determine their overall efficacy at communicating uncertainty in the context of AP Statistics exam questions. Unlike Study 1, this research was conducted online in order to ensure that it would be accessible to all participants regardless of sensory ability. The auditory displays were assessed both on their own and together with both error bars and the alternative display created for Study 1. There was no visual only condition because this would

prevent the study from being fully accessible. As with Study 1, confidence was also assessed to determine what, if any, role it played in overall performance.

### *7.2.1. Participants*

Participants for this study were 66 members of Georgia Tech and the surrounding community (54 sighted and 12 with vision loss), recruited through the university SONA system or through word of mouth. Participants recruited through SONA received 1 credit per hour spent completing the study. Participants had normal or corrected-to-normal hearing and completed either a first semester statistics course at an accredited university and/or the statistics AP exam.

### *7.2.2. Experiment Structure*

While Study 1 focused on the evaluation of visual displays of confidence intervals and their effect on student performance, Study 2 focused on auditory representations and focuses on answering the final proposed research question. Like the prior study, participants were asked to complete an abbreviated version of the free response section of the AP Statistics exam, but in this case the exam was administered digitally using Qualtrics rather than paper and pencil. The reason for this is to ensure that the experiment is fully accessible to visually impaired participants that take part in the study. In addition, it also allowed for the easier integration of auditory displays into the exam/question structure. Participant confidence data as well as training types were presented and collected in the same manner as they were in experiment 1.

In this study, participants again completed 6 questions, but all of them were on confidence intervals. Three of the questions utilized auditory uncertainty displays crafted with first order filters and three used displays made with second order filters. For each of these display types, participants got one question that is audio only, one that is audio with visual

confidence intervals, and one that is audio plus an alternative visual uncertainty display.

### 7.2.3. Question Selection

Questions were selected in the same manner as experiment 1.

### 7.2.4. Auditory Stimulus Creation

The auditory uncertainty displays used in this study were created based on scaling equations calculated in Batterman & Walker (2012a, 2012b, 2013) and created using Audacity.

The equations used to calculate the tempos and frequencies are shown below:

Tempo:

$$y = 0.5361x^{0.6669}$$

$$FV = x \bullet 2.369833$$

Frequency:

$$y = 0.7996x^{0.452}$$

$$FV = x \bullet 3.621566$$

The first equation in each pair was taken from Batterman & Walker (2012b) and is the scaling equation for tempo and frequency respectively. The distance between the bounds of the confidence interval is plugged in as  $y$  in the equations and solving for  $x$  would yield the tempo in bpm or frequency in Hz respectively. However, the  $x$  values generated for any  $y$  value 8 or below are outside the ranges that the scaling equations were validated with in Batterman & Walker (2013) (60-300 bpm and 200-1000 Hz). Because this is the first application of these mappings in this context, constraining our values between a subset of already validated frequency and tempo ranges is the best way to ensure the display is successful at conveying the desired information. To do this, the second equation in each pair was created by adapting an

equation from Batterman & Walker (2013):

$$FV = IV \bullet (S_5/S_1)^m$$

This equation was created to calculate the ‘Final Value’ (FV) of a numeric pair that was set to match up to the uncertainty level represented by a sequence of ascending frequencies or tempos. The final value in an ascending set was  $S_5$  and the initial value was  $S_1$ , because  $300/60$  and  $1000/200$  both equal 5, that was plugged in in their place for the purposes of this study. The  $m$  represents the slope of the scaling equation, so the slopes for the initial scaling equations were plugged in there. That yielded values of 2.369833 for tempo and 3.621566 for frequency.  $IV$ , which stood for initial value in the original equation was replaced with the  $x$  calculated from the first equation, to yield the completed second equation.

Once the final values were calculated for a given question, the auditory displays could be created. Each display consisted of a brief tone (created by typing the mean values into the Sonification Sandbox) being played representing the mean, followed by 2 seconds of filtered noise pulsing at the bpm calculated by the first equation set. The frequency value became the central frequency for a bandpass filter created over white noise. Two versions of each display were created, one with a first order filter with a 6 dB decay and one with a second order filter with a 12 dB decay.

#### *7.2.5. Supplemental Materials Creation*

Because the materials provided for the AP Statistics exam were not accessible, accessible materials needed to be created and provided to participants so that those who were visually impaired would have access to the same information as their sighted counterparts. To create accessible  $z$ ,  $t$ , and chi-squared tables, the values were entered into and tagged in Microsoft Excel and provided as downloadable files to participants. To recreate the equation sheet, a

downloadable HTML file was created rendering each of the equations in MathML, along with instructions on how to download MathPlayer and MathJax to ensure that the equations rendered properly.

#### *7.2.6. Procedure*

Prior to their completing the study, participants filled out their basic demographics information online, in addition to affirming that they meet the criterion to participate in the study. They were then randomly assigned to either receive confidence interval training or probability training.

Following this they were sent a link to the Qualtrics survey, which guided them through a consent form and a list of approved calculators they could use during the assessment. At this point they were presented with the confidence measures to fill out briefly before moving on to their training. At the end of the training, they were again asked to fill out these confidence measures. Following this they moved on to the exam questions and were told they had one hour to complete them. When they had completed the questions, they were debriefed and told they could close their browsers.

#### *7.2.7. Hypotheses/Expected Results*

The proposed study evaluated the effect that different display types have on confidence interval questions and whether or not this is mitigated at all by training and exposure to them. It also assessed the confidence participants have in their own abilities, whether training can increase this, and if this increase correlates with superior performance. Furthermore, it also assesses the evaluation of auditory uncertainty as displayed by two different filters, determines the extent that auditory uncertainty displays are effected by training (especially when compared to the results of Study 1 with visual displays), and determines the extent that auditory displays

can work in congress with visual displays to effect performance.

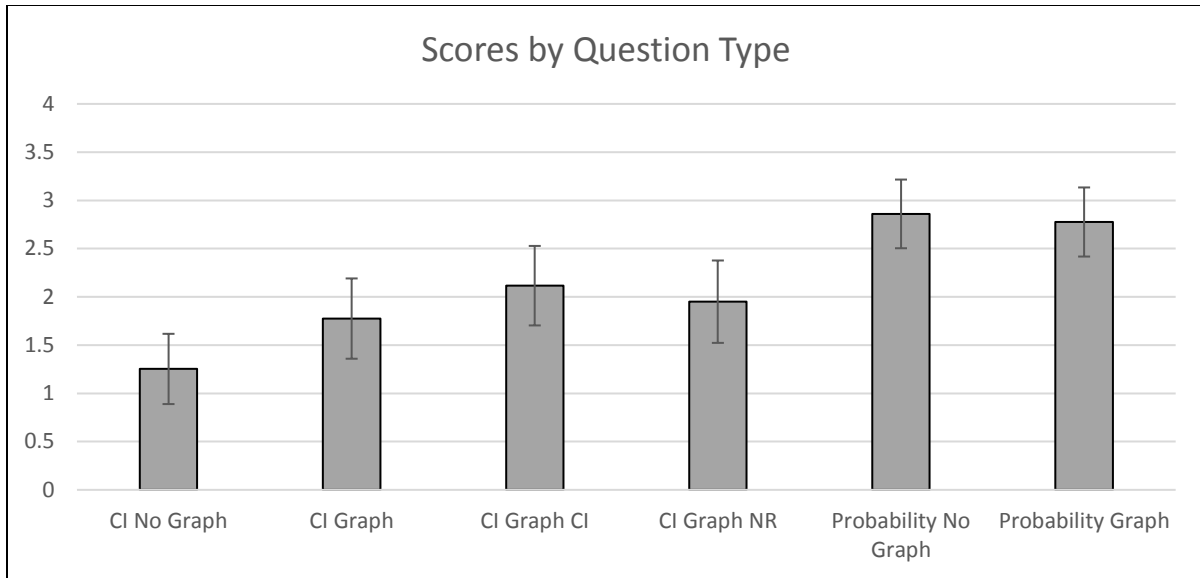
## CHAPTER 8. DATA ANALYSIS AND RESULTS

### 8.1. Study 1 Analysis

The primary analysis was a 2 (confidence interval training vs probability training) x 6 (confidence interval questions, with graph, with graph and confidence interval display, and with alternative display and two questions on other statistical topics with and without graphs) mixed measures ANOVA with training being a between-subjects variable and question/display type being within-subjects. Bonferroni corrected paired t tests were conducted post hoc on the question/display type factor to determine potential differences. The difference scores calculated from the pre- and post-training confidence measures was included in the omnibus analysis as covariates to determine the effect any potential change in confidence levels would have on performance. In addition, an independent samples t test was conducted to see if training had an overall effect on confidence regardless of whether or not it impacted performance. Other factors such as number of statistics courses taken, whether participants took the AP Statistics exam, and their grades in statistics courses were also taken into account and included in subsequent analyses to determine their effects.

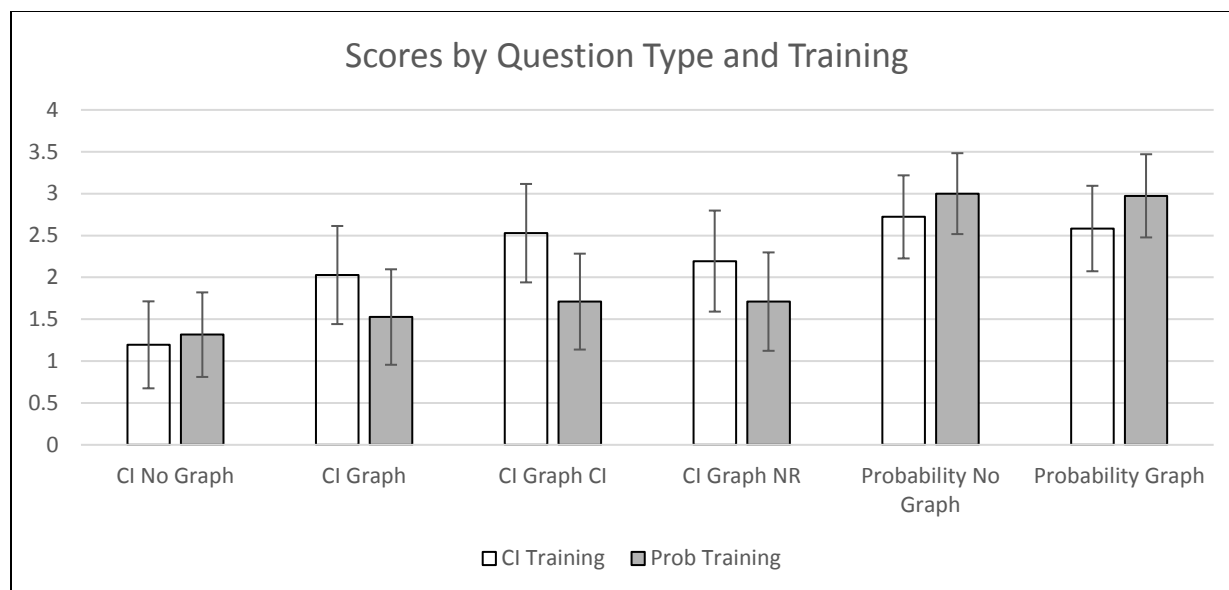
#### 8.1.1. Study 1 Results

A 2x6 mixed measures ANOVA was conducted with confidence as a covariate and revealed a main effect of Question Type  $F(4.457, 156) = 19.205, p < 0.05, \eta_p^2 = 0.354$  and a significant interaction between question type and condition  $F(4.457, 156) = 3.139, p < 0.05, \eta_p^2 = 0.082$  (both corrected with Huynh-Feldt) (see Figure 12). Planned post hocs (Bonferroni corrected paired t-tests) were conducted between question types and display factors to both explain the main effect and parse out the interaction. For the confidence interval questions,



**Figure 12 - Bar graph depicting differences in scores between question types collapsed across training condition. Error bars represent 0.95 confidence intervals.**

every condition with a visual display performed significantly greater than their no display counterpart (CI Graphs vs. No Graphs:  $t(36) = 2.954, p < 0.05$ , CI Graphs with Error Bars vs. No Graphs:  $t(36) = 4.155, p < 0.05$ , CI Graphs with New Representation vs. No Graphs:  $t(36) = 5.013, p < 0.05$ ). Furthermore, confidence interval questions with graphs and error bars scored significantly higher than those with graphs alone  $t(36) = 1.913, p < 0.05$ , but this was not the case for the new representation  $t(36) = 1.039, p > 0.05$  (see Figure 13). Unlike the confidence interval condition, probability questions did not differ by display type  $t(36) = 0.420, p > 0.05$ . In addition, confidence was not a significant covariate in the overall performance analysis (Math



**Figure 13 - Bar graph depicting differences in scores between split by both question type and training condition. Error bars represent 0.95 confidence intervals.**

Confidence:  $F(1, 33) = 1.586, p > 0.05$ ; Statistics Confidence:  $F(1, 33) = 0.007, p > 0.05$ .

However, math confidence was significantly different as a result of training condition, with confidence levels being higher after being exposed to probability training than they were after being exposed to confidence interval training,  $t(35) = 2.518, p < 0.05$ . Statistics confidence was not found to be significant across the same parameters,  $t(35) = 1.079, p > 0.05$ . Participants' confidence in themselves as test takers was also found to not differ across conditions,  $t(35) = 0.812, p > 0.05$ .

A series of additional tests were conducted to determine if participants' experience in college statistics courses, their grades, whether they had taken the AP Statistics course or exam, or their subsequent score had any overall effect on their performance. To determine this, a MANOVA was conducted with scores on the various question types as dependent variables, whether participants had taken the AP exam, AP class, and/or college statistics as fixed factors,

and their scores on these dimensions as covariates. Results demonstrated a significant relationship between students' scores on the AP exam and their scores on the two probability questions, with those who scored higher on the exam performing significantly better on those questions (Probability No Graphs,  $F(1, 7) = 7.738, p < 0.05, \eta_p^2 = 0.607$  and Probability Graphs,  $F(1, 7) = 10.06, p < 0.05, \eta_p^2 = 0.668$ ).

### *8.1.2. Study 1 Data Explained*

Taken as a whole, the results of this study confirm many of the initial hypotheses while answering many of the research questions as well. Each of the relevant research questions are addressed individually, below, followed by a summary of findings.

*R1: Can confidence intervals be better understood with alternative visual display types?*

*Answer: Inconclusive*

The new representation performed comparably to traditional error bars in every instance except for in the confidence interval training condition, where questions containing graphs with error bars scored significantly higher than those without them, but the same was not true for the new representation. This is likely due to both prior exposure to error bars as well as more developed training scripts. Because this representation was created for the purpose of this study, the training script was developed from scratch rather than being drawn from existing education literature. Furthermore, though graphs with the new representation failed to reach significance when compared to graphs with no uncertainty representation, they were also not statistically different from error bars. This suggests that both the new representation and the traditional error bar are equally effective. Though the representation created for this study did not perform better than an error bar, its comparable performance suggests that alternative representations are worth

exploring to truly maximize the potential of visual displays for conveying uncertainty.

*R2: Do questions involving confidence intervals have a unique status when it comes to the extent they are better understood with graphs present?*

*Answer: Yes*

Not only did confidence interval questions consistently score worse than probability questions, but performance on those questions increased with graphs present whereas that was not the case with probability questions. This suggests that for the concept of uncertainty, graphs hold a unique ability to enhance performance above and beyond other comparably difficult concepts like probability.

*R3: Can the presence of these graphs enhance performance even further if uncertainty is also represented on them?*

*Answer: Yes*

Confidence interval questions with graphs and error bars scored significantly higher than those with graphs alone, which in turn scored higher than those with no representation. This demonstrates the importance of uncertainty representations on facilitating understanding of the underlying concepts. It also stands contrary to views which suggest that error bars should not be shown because they are not well understood. The results of this study provide evidence that the presence of error bars enhances performance overall beyond that of visuals with without them,

and thus they should be added to displays whenever the relevant information exists to create them.

*R4: Can confidence and performance be increased with prior training and explanation?*

*Answer: Yes*

There was a significant interaction between training (confidence interval vs. probability) and question type, with probability questions scoring higher with probability training and confidence interval questions with displays scoring higher in the confidence interval condition. This is important because it demonstrates that without prior explanation, confidence interval displays still score higher than questions without them, but that this gap widens with training. It furthers the case that these displays should be essential learning in any statistics course.

Confidence in one's own mathematics ability was shown to increase in the probability condition, but this increase was not shown to affect overall performance. The confidence increase is likely due to participants being presented with information they already had some understanding of (they all had had some prior statistics training) and that bolstering their overall confidence. However, despite being a statistical topic, confidence in statistics abilities did not increase, this may be due to probability not being perceived as a purely statistical topic because it appears in many branches of mathematics.

### *8.1.3. Study 1 Summary Discussion*

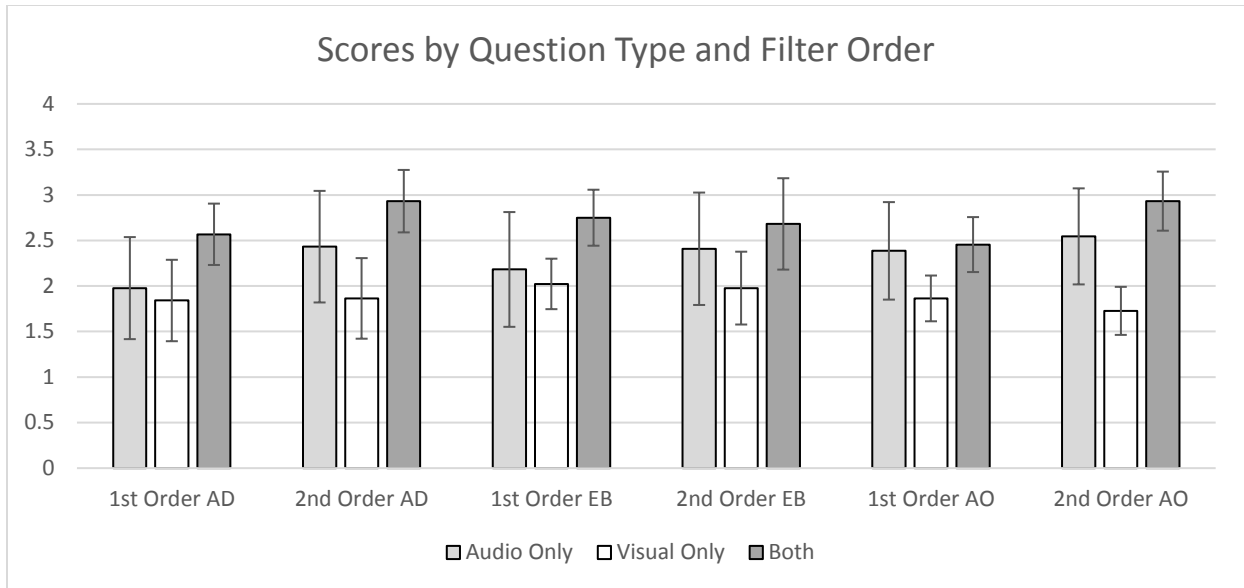
As a whole Study 1 was successful in demonstrating the importance of the error bar representation in peoples' understanding of confidence intervals as a whole. This understanding was further bolstered due to the brief training participants received. This offers strong evidence

for not only continuing to use error bars, but also for teaching and including them in statistics courses as a whole.

Though the new representation failed to outperform error bars as a measure of uncertainty, its' performance was comparable and it certainly merits more study. It also demonstrated that uncertainty representations are so important, that even a novel one that participants had never encountered before (albeit created to be seen as more uncertain) added to overall understanding and performance.

## **8.2. Study 2 Analyses**

The primary analysis was a 3 (visual confidence interval training, auditory confidence interval training, both) x 2 (1<sup>st</sup> order filters, 2<sup>nd</sup> order filters) x 3 (auditory display only, auditory display with visual error bar, auditory display with alternative confidence interval display) ANOVA with the first factor being between-subjects and the rest of the study being within-subjects. Orthogonal Bonferroni corrected paired t-tests will be conducted on the display and training dimensions and self- confidence measures will be added to the analysis as covariates. Other factors such as number of statistics courses taken, whether participants took the AP



**Figure 14 - Bar graph depicting differences in scores between training conditions for both 1<sup>st</sup> and 2<sup>nd</sup> order filters. Error bars represent 0.95 confidence intervals.**

Statistics exam, and their grades in statistics courses were also taken into account and included in subsequent analyses to determine their effects.

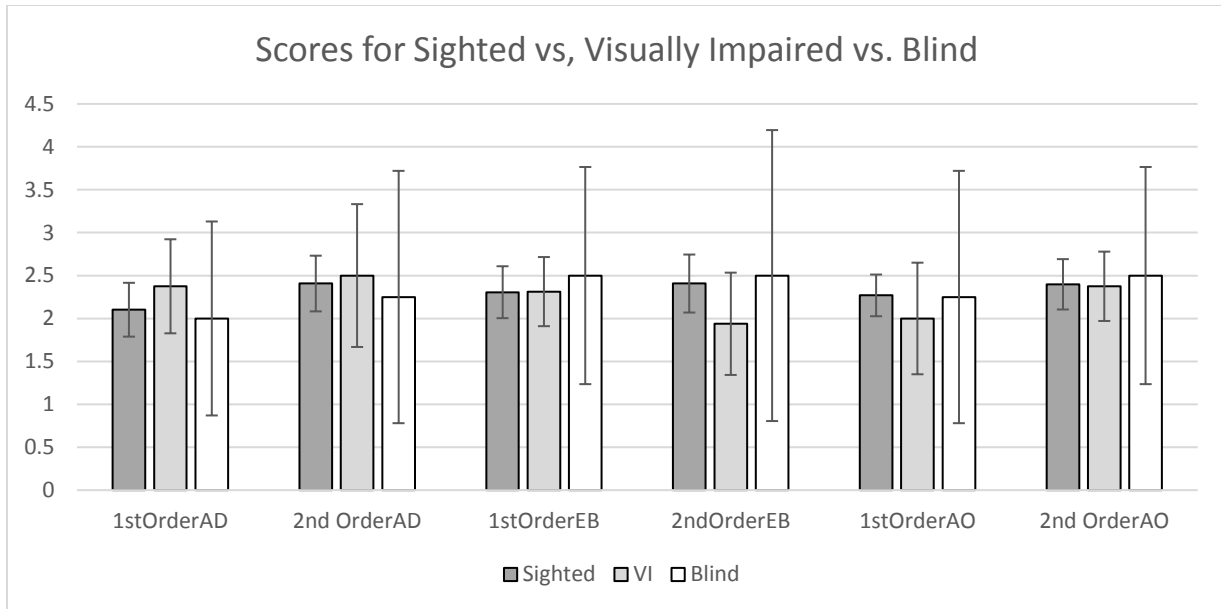
### 8.2.1. Study 2 Results

The results of the 3 (visual confidence interval training, auditory confidence interval training, both) x 2 (1<sup>st</sup> order filters, 2<sup>nd</sup> order filters) x 3 (auditory display only, auditory display with visual error bar, auditory display with alternative confidence interval display) mixed measures ANOVA with confidence measures as covariates revealed a main effect of training condition,  $F(2, 63) = 5.068, p < 0.05, \eta_p^2 = 0.139$ , and filter,  $F(1, 126) = 6.506, p < 0.05, \eta_p^2 = 0.094$ , but no other significant main effects or interactions (see Figure 14). Planned post hocs (Bonferroni corrected paired t-tests) revealed a significant difference ( $p < 0.05$ ) between the visual only and both conditions, with the latter having higher scores. In addition, confidence was not a significant covariate in the overall performance analysis (Math Confidence:  $F(1, 61) =$

0.003,  $p > 0.05$ ; Statistics Confidence:  $F(1, 61) = 1.864, p > 0.05$ ).

However, statistics confidence was significantly different as a result of training condition, with confidence levels being higher after being exposed to both training conditions,  $t(21) = 3.196, p < 0.05$ , than then they were after being exposed to visual or audio training alone. Math confidence was not shown to be different across conditions. Participants' confidence in themselves as test takers was also found to not differ across conditions,  $t(21) = 0.413, p > 0.05$ .

As with Study 1, a series of additional tests were conducted to determine if participants' experience in college statistics courses, their grades, whether they had taken the AP Statistics course or exam, or their subsequent score had any overall effect on their performance. To determine this, a MANOVA was conducted with scores on the questions as dependent variables, whether or not participants has taken the AP exam or college statistics as fixed factors, and their scores on the AP exam and college courses as covariates. Results demonstrated a significant relationship between scores on the AP exam and performance on questions that used second order filters and had an audio display only,  $F(1, 10) = 5.336, p < 0.05, \eta_p^2 = 0.508$ , but no other significant effects. Further analysis, demonstrates that the majority of participants who took the AP exam in the study scored a 4 or a 5 (18 out of the 21 participants), and had a mean score for second order filter audio display questions of 2.528. Those 3 participants who scored lower than a 4 averaged scores of 1.33 on the same questions. This is likely the explanation for the significant effect observed in the omnibus. Because there were so few participants who took the AP exam and scored below a 4, it is tough to generalize much from this result. It is also unknown why this effect was just observed for just questions with audio displays and second



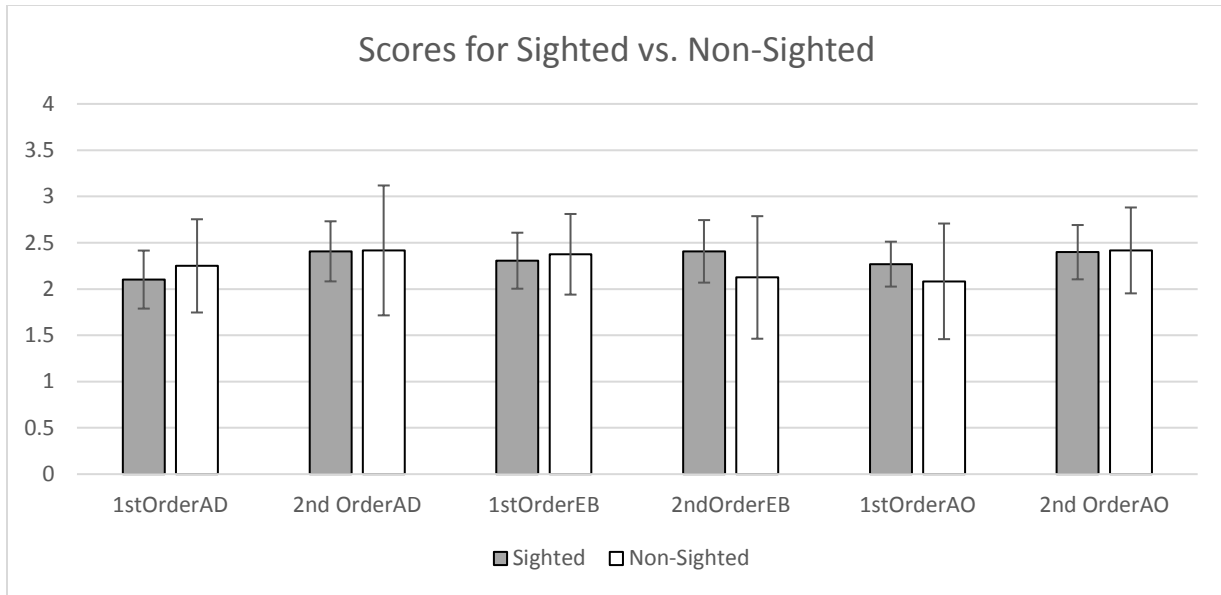
**Figure 15 - Bar graph depicting differences in scores between conditions for participants who identified as sighted, blind, and visually impaired. Error bars represent 0.95 confidence intervals.**

order filters, but it is likely an artifact of the small sample size of participants in this group.

Of the 66 participants in Study 2, 54 were sighted, 8 identified as visually impaired, and 4 as blind. No significant performance difference was found between these groups both when individuals with vision impairment and those that identified as blind were analyzed together, or separately (see Figures 15 and 16). There was also no significant difference in confidence between these groups.

### 8.2.2. Study 2 Data Explained

Taken as a whole, the results of this study are extremely promising. Not only did it demonstrate the effectiveness of auditory representations of uncertainty, but it also demonstrated the



**Figure 16 - Bar graph depicting differences in scores between conditions for participants who identified as sighted vs. those who did not. Error bars represent 0.95 confidence intervals.**

increased benefit of using both visual and auditory representations together. Participants who were trained in both conditions not only scored higher, but were more confident in their statistics abilities as well. This is a prime example of the importance of presenting information in different ways across different modalities. Next the pertinent research questions will be addressed, followed by a summary of the findings as a whole.

*R4: Can confidence and performance be increased with prior training and explanation?*

Like with Study 1, the answer to this question is yes. When participants received prior training with both audio and visual displays together, their overall performance increased above and beyond when they were trained on either modality on its own. Furthermore, this cross-modality training increased their confidence. This reinforces the findings of study 1, further demonstrating that confidence intervals and uncertainty displays in general should be taught and

explained to enhance understanding. It adds to this by offering evidence that a multimodal approach should be taken when teaching the concept to not only ensure accessibility, but also to increase the overall utility of the display itself.

*R5: Can use of other modalities [i.e. auditory displays] enhance student performance in addition to making the displays accessible?*

In short, the answer to this is a resounding yes. This is the first study to create and deploy an auditory uncertainty measure in the context of a testing environment, both as a substitute for and/or an enhancement to an existing visual uncertainty display. In terms of enhancement, the presence of auditory displays when combined with the visual yielded a significant boost in both performance and confidence. Furthermore, participants with varying degrees of vision impairment and equivalent statistical backgrounds performed equal to their sighted counterparts. This is an important finding in that it makes these displays viable and deployable in an educational context. One of the more difficult aspects of designing for this space is ensuring that all students are receiving equivalent information without any student being afforded an unfair advantage. The sort of display used in this study is not only accessible, but was experimentally demonstrated to be equivalent in a facsimile of a testing environment.

### *8.2.3. Study 2 Summary Discussion*

As a whole, this study serves to not only demonstrate that a viable and functionally equivalent auditory uncertainty display can be developed, but that it can benefit students of all abilities by enhancing performance above and beyond the existing visual displays. Furthermore, the increase in confidence is important, not only because confidence is important as a whole, but also because it affected participants with and without visual impairments equally. If anything, it can be argued that students with vision loss need confidence just as much as they need effective

displays so that they can hopefully leverage it into the pursuit of careers in STEM, a grouping of fields that is currently lacking the unique diversity and insight that many without vision can provide. As an aside, it took around 6 months to recruit even the 12 participants with vision loss that participated in this study. Finding someone with vision loss and the requisite statistics experience was extremely difficult due to the rarity of someone in this group who had completed even one college level statistics class. These students consistently face barriers from both instructors and uninformed friends and family members who convince them that mathematics that rely heavily on visual displays are something they cannot do. With these findings from this study, there is hope that a piece of that puzzle will become more accessible, and thus open up a new world to potential blind students.

## CHAPTER 9. GENERAL DISCUSSION AND CONCLUSIONS

Taken together, these studies answer important questions about how uncertainty is represented and processed, as well as how modality and display type effect it. Results of Study 1 demonstrate the importance of both displaying uncertainty as a whole, and educating people on how it is represented; while results of Study 2 show the benefits of auditory uncertainty displays for both sighted and visually impaired users (with the latter having access to a functionally equivalent representation for the first time).

For questions involving uncertainty, the presence of a visual display alone increased performance, but this increase was even greater when an uncertainty display was present. This difference was magnified when participants were given some training on uncertainty displays and how to interpret them. This finding suggests that rather than ‘getting rid’ of error bars as some have suggested, the focus instead should be on teaching people how to use and interpret them.

Though the new uncertainty display created for this study did not outperform error bar representations, they did match them. Though this did not live up to performance expectations, it is still a promising sign. Regardless of whether or not they knew how to interpret them or not, participants had likely seen or been exposed to error bars before at some point during their life and/or education and the same could not be said of the new representation, which they were seeing for the first time. The fact that performance was at least on par with existing representations is an excellent first step, and future iterations of the display will endeavor to improve performance even further.

Auditory uncertainty displays were also shown to be viable ways to convey uncertainty,

both with and without visual displays present. The fact that this was shown in a testing context is even more powerful in that it demonstrates how these displays could be deployed to give visually impaired students access to functionally equivalent displays in a testing environment; displays that are now either described textually or produced tactilely at much higher cost. It also moves closer to universally designing the digital testing environment as a whole. If tests are administered digitally, providing the graphical information (including uncertainty) to all test takers in multiple modalities would not only make the tests more accessible, but also likely bolster performance as students would benefit from being exposed to many display types.

Furthermore, though the effect size was small, the significant difference in performance between audio created with first and second order filters lends credence to the importance of creating psychologically salient displays. The additional resonance provided by the second order filters, however subtle, added an additional element that was previously identified as uncertain, bolstering the overall strength of the display. This is important for uncertainty, auditory display creation, and display creation as a whole as it shows how important it is that displays are designed with their intended meaning in mind.

Future work should focus on refining both the auditory and visual displays created for this series of studies to make them even more effective and to expand their usage to different and broader contexts. Though it matched standard error bars in this context, it may exceed them in broader uncertainty representations outside a standard graph. In addition, the display itself may need to be altered to further stress the inherent uncertainty that it is intended to convey. For the auditory display, it should be refined further as well, but its exceptional performance in this context (especially when coupled with a visual display) suggests that it should be ready to be tested and integrated into other software as well, potentially conveying uncertainty about

geographical locations, medical scans, and more advanced mathematical simulations and models.

Finally, the results of this work stress how important it is to expose students to uncertainty displays and teach them how to use and interpret them. The lack of focus paid to them in introductory statistics courses (AP and otherwise) is frankly baffling. The brief exposure in this study increased performance significantly and it is very likely that if these displays were taught formally, it could go up even more. Uncertainty displays seem to have a unique status related to performance in that they cause it to increase in ways that are not seen in other dimensions (i.e., probability). It is almost as if, ironically enough, they are able to increase the clarity of the information being displayed so that it can be better understood. Education and prior exposure can (and do) increase performance, but only if their importance is stressed and the information is conveyed consistently and correctly. The analyses from these studies demonstrated that students who had previously taken the AP exam performed significantly better on the probability questions than those who had not taken the exam, but this did not happen with uncertainty questions. This suggests that these students not only focused more on probability questions, but that they learned and understood it better. Furthermore, that uncertainty has a unique status when it comes to understanding being bolstered by graphs/displays, suggesting that these should be taught along with the rote concepts to increase understanding.

Taken together these studies contribute to bodies of work in the fields of visual display, auditory display, educational testing, and psychophysics and lay the groundwork for better uncertainty displays to be created and deployed in the future.

## Appendix A

### Questions for Experiments

#### Appendix A1

#### Confidence Interval Questions with No Graphs

##### *Question 4 from 2006 Exam*

Patients with heart-attack symptoms arrive at an emergency room either by ambulance or self-transportation provided by themselves, family, or friends. When a patient arrives at the emergency room, the time of arrival is recorded. The time when the patient's diagnostic treatment begins is also recorded.

An administrator of a large hospital wanted to determine whether the mean wait time (time between arrival and diagnostic treatment) for patients with heart-attack symptoms differs according to the mode of transportation. A random sample of 150 patients with heart-attack symptoms who had reported to the emergency room was selected. For each patient, the mode of transportation and wait time were recorded. Summary statistics for each mode of transportation are shown below.

##### Mode of Transportation

Ambulance, Sample Size = 77    Mean Wait Time (in minutes) = 6.04    Standard Deviation of Wait Times (in minutes) = 4.30  
Self, Sample Size = 73    Mean Wait Time (in minutes) = 8.30    Standard Deviation of Wait Times (in minutes) = 5.16

- (a) Use a 99 percent confidence interval to estimate the difference between the mean wait times for ambulance-transported patients and self-transported patients at this emergency room.
- (b) Based only on this confidence interval, do you think the difference in the mean wait times is statistically significant? Justify your answer.

Question 1 from 2007 Exam

**Directions:** Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. The department of agriculture at a university was interested in determining whether a preservative was effective in reducing discoloration in frozen strawberries. A sample of 50 ripe strawberries was prepared for freezing. Then the sample was randomly divided into two groups of 25 strawberries each. Each strawberry was placed into a small plastic bag.

The 25 bags in the control group were sealed. The preservative was added to the 25 bags containing strawberries in the treatment group, and then those bags were sealed. All bags were stored at 0°C for a period of 6 months. At the end of this time, after the strawberries were thawed, a technician rated each strawberry's discoloration from 1 to 10, with a low score indicating little discoloration.

The data below show the distributions of discoloration rating for the control and treatment groups.

Treatment

Rating	1	2	3	4	5	6	7	8	9	10
Number	2	1	3	3	6	4	2	1	2	1

Control

Rating	1	2	3	4	5	6	7	8	9	10
Number	0	1	2	1	2	6	4	4	3	2

- (a) The standard deviation of ratings for the control group is 2.141. Explain how this value summarizes variability in the control group.
- (b) Based on the raw data, comment on the effectiveness of the preservative in lowering the amount of discoloration in strawberries. (No calculations are necessary.)
- (c) Researchers at the university decided to calculate a 95 percent confidence interval for the difference in mean discoloration rating between strawberries that were not treated with preservative and those that were treated with preservative. The confidence interval they obtained was (0.16, 2.72). Assume that the conditions necessary for the  $t$ -confidence interval are met.

Based on the confidence interval, comment on whether there would be a difference in the population mean discoloration ratings for the treated and untreated strawberries.

*Question 4 from 2009 Exam*

One of the two fire stations in a certain town responds to calls in the northern half of the town, and the other fire station responds to calls in the southern half of the town. One of the town council members believes that the two fire stations have different mean response times. Response time is measured by the difference between the time an emergency call comes into the fire station and the time the first fire truck arrives at the scene of the fire.

Data were collected to investigate whether the council member's belief is correct. A random sample of 50 calls selected from the northern fire station had a mean response time of 4.3 minutes with a standard deviation of 3.7 minutes. A random sample of 50 calls selected from the southern fire station had a mean response time of 5.3 minutes with a standard deviation of 3.2 minutes.

- (a) Construct and interpret a 95 percent confidence interval for the difference in mean response times between the two fire stations.
- (b) Does the confidence interval in part (a) support the council member's belief that the two fire stations have different mean response times? Explain.

Question 5 from 2016 Exam

A polling agency showed the following two statements to a random sample of 1,048 adults in the United States.

Environment statement: Protection of the environment should be given priority over economic growth.

Economy statement: Economic growth should be given priority over protection of the environment.

The order in which the statements were shown was randomly selected for each person in the sample. After reading the statements, each person was asked to choose the statement that was most consistent with his or her opinion. The results are shown in the table.

	Environment Statement	Economy Statement	No Preference
Percent of sample	58%	37%	5%

- Assume the conditions for inference have been met. Construct and interpret a 95 percent confidence interval for the proportion of all adults in the United States who would have chosen the economy statement.
- One of the conditions for inference that was met is that the number who chose the economy statement and the number who did not choose the economy statement are both greater than 10. Explain why it is necessary to satisfy that condition.
- A suggestion was made to use a two-sample  $z$ -interval for a difference between proportions to investigate whether the difference in proportions between adults in the United States who would have chosen the environment statement and adults in the United States who would have chosen the economy statement is statistically significant. Is the two-sample  $z$ -interval for a difference between proportions an appropriate procedure to investigate the difference? Justify your answer.

## Appendix A2

### Confidence Interval Questions with Graphs

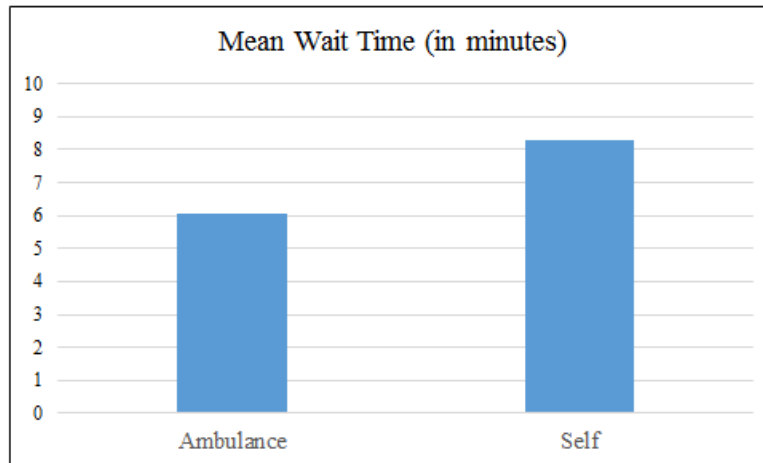
#### Question 4 from 2006 Exam

Patients with heart-attack symptoms arrive at an emergency room either by ambulance or self-transportation provided by themselves, family, or friends. When a patient arrives at the emergency room, the time of arrival is recorded. The time when the patient's diagnostic treatment begins is also recorded.

An administrator of a large hospital wanted to determine whether the mean wait time (time between arrival and diagnostic treatment) for patients with heart-attack symptoms differs according to the mode of transportation. A random sample of 150 patients with heart-attack symptoms who had reported to the emergency room was selected. For each patient, the mode of transportation and wait time were recorded. Summary statistics for each mode of transportation are shown below.

#### Mode of Transportation

Ambulance, Sample Size = 77    Mean Wait Time (in minutes) = 6.04    Standard Deviation of Wait Times (in minutes) = 4.30  
Self, Sample Size = 73    Mean Wait Time (in minutes) = 8.30    Standard Deviation of Wait Times (in minutes) = 5.16



- Use a 99 percent confidence interval to estimate the difference between the mean wait times for ambulance-transported patients and self-transported patients at this emergency room.
- Based only on this confidence interval, do you think the difference in the mean wait times is statistically significant? Justify your answer.

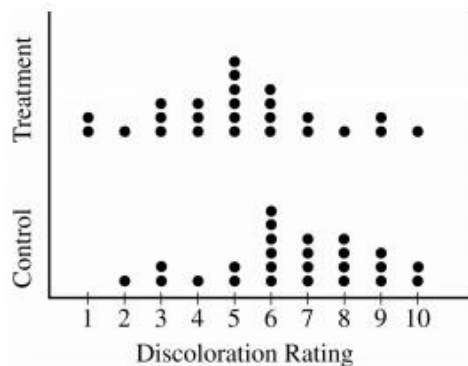
Question 1 from 2007 Exam

**Directions:** Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. The department of agriculture at a university was interested in determining whether a preservative was effective in reducing discoloration in frozen strawberries. A sample of 50 ripe strawberries was prepared for freezing. Then the sample was randomly divided into two groups of 25 strawberries each. Each strawberry was placed into a small plastic bag.

The 25 bags in the control group were sealed. The preservative was added to the 25 bags containing strawberries in the treatment group, and then those bags were sealed. All bags were stored at  $0^{\circ}\text{C}$  for a period of 6 months. At the end of this time, after the strawberries were thawed, a technician rated each strawberry's discoloration from 1 to 10, with a low score indicating little discoloration.

The dotplots below show the distributions of discoloration rating for the control and treatment groups.



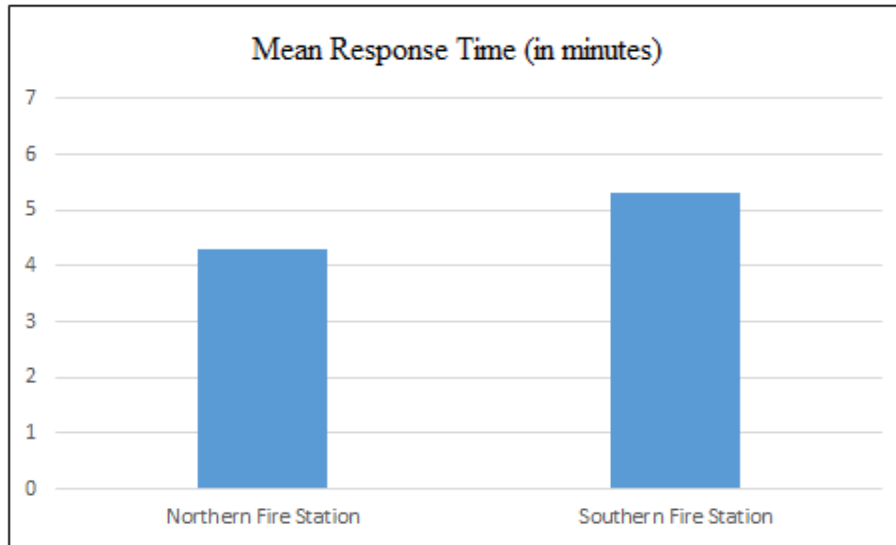
- (a) The standard deviation of ratings for the control group is 2.141. Explain how this value summarizes variability in the control group.
- (b) Based on the dotplots, comment on the effectiveness of the preservative in lowering the amount of discoloration in strawberries. (No calculations are necessary.)
- (c) Researchers at the university decided to calculate a 95 percent confidence interval for the difference in mean discoloration rating between strawberries that were not treated with preservative and those that were treated with preservative. The confidence interval they obtained was (0.16, 2.72). Assume that the conditions necessary for the  $t$ -confidence interval are met.

Based on the confidence interval, comment on whether there would be a difference in the population mean discoloration ratings for the treated and untreated strawberries.

*Question 4 from 2009 Exam*

One of the two fire stations in a certain town responds to calls in the northern half of the town, and the other fire station responds to calls in the southern half of the town. One of the town council members believes that the two fire stations have different mean response times. Response time is measured by the difference between the time an emergency call comes into the fire station and the time the first fire truck arrives at the scene of the fire.

Data were collected to investigate whether the council member's belief is correct. A random sample of 50 calls selected from the northern fire station had a mean response time of 4.3 minutes with a standard deviation of 3.7 minutes. A random sample of 50 calls selected from the southern fire station had a mean response time of 5.3 minutes with a standard deviation of 3.2 minutes.



- (a) Construct and interpret a 95 percent confidence interval for the difference in mean response times between the two fire stations.
- (b) Does the confidence interval in part (a) support the council member's belief that the two fire stations have different mean response times? Explain.

Question 5 from 2016 Exam

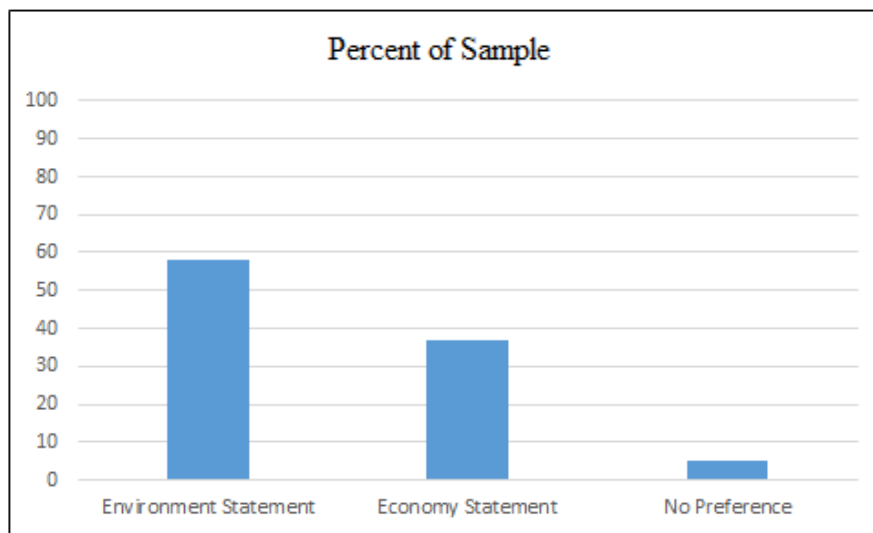
A polling agency showed the following two statements to a random sample of 1,048 adults in the United States.

Environment statement: Protection of the environment should be given priority over economic growth.

Economy statement: Economic growth should be given priority over protection of the environment.

The order in which the statements were shown was randomly selected for each person in the sample. After reading the statements, each person was asked to choose the statement that was most consistent with his or her opinion. The results are shown in the table.

	Environment Statement	Economy Statement	No Preference
Percent of sample	58%	37%	5%



- Assume the conditions for inference have been met. Construct and interpret a 95 percent confidence interval for the proportion of all adults in the United States who would have chosen the economy statement.
- One of the conditions for inference that was met is that the number who chose the economy statement and the number who did not choose the economy statement are both greater than 10. Explain why it is necessary to satisfy that condition.
- A suggestion was made to use a two-sample  $z$ -interval for a difference between proportions to investigate whether the difference in proportions between adults in the United States who would have chosen the environment statement and adults in the United States who would have chosen the economy statement is statistically significant. Is the two-sample  $z$ -interval for a difference between proportions an appropriate procedure to investigate the difference? Justify your answer.

## Appendix A3

### Confidence Interval Questions with Traditional Representation (Error Bars)

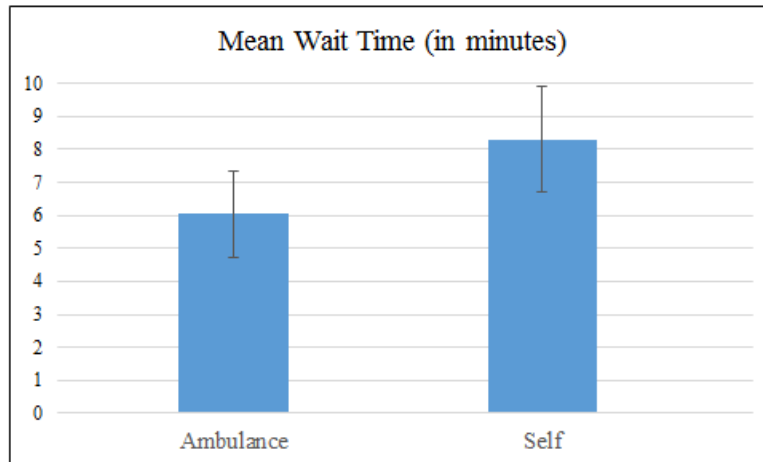
#### *Question 4 from 2006 Exam*

Patients with heart-attack symptoms arrive at an emergency room either by ambulance or self-transportation provided by themselves, family, or friends. When a patient arrives at the emergency room, the time of arrival is recorded. The time when the patient's diagnostic treatment begins is also recorded.

An administrator of a large hospital wanted to determine whether the mean wait time (time between arrival and diagnostic treatment) for patients with heart-attack symptoms differs according to the mode of transportation. A random sample of 150 patients with heart-attack symptoms who had reported to the emergency room was selected. For each patient, the mode of transportation and wait time were recorded. Summary statistics for each mode of transportation are shown below.

#### Mode of Transportation

Ambulance, Sample Size = 77    Mean Wait Time (in minutes) = 6.04    Standard Deviation of Wait Times (in minutes) = 4.30  
Self, Sample Size = 73    Mean Wait Time (in minutes) = 8.30    Standard Deviation of Wait Times (in minutes) = 5.16



- Use a 99 percent confidence interval to estimate the difference between the mean wait times for ambulance-transported patients and self-transported patients at this emergency room.
- Based only on this confidence interval, do you think the difference in the mean wait times is statistically significant? Justify your answer.

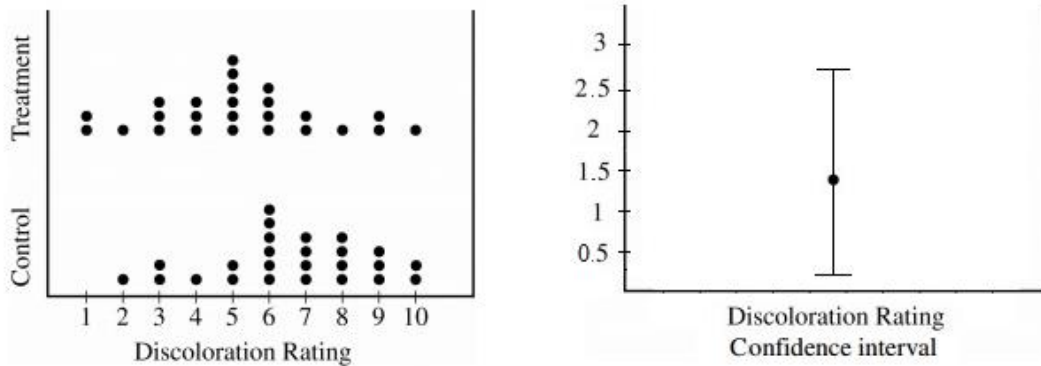
Question 1 from 2007 Exam

**Directions:** Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

- The department of agriculture at a university was interested in determining whether a preservative was effective in reducing discoloration in frozen strawberries. A sample of 50 ripe strawberries was prepared for freezing. Then the sample was randomly divided into two groups of 25 strawberries each. Each strawberry was placed into a small plastic bag.

The 25 bags in the control group were sealed. The preservative was added to the 25 bags containing strawberries in the treatment group, and then those bags were sealed. All bags were stored at  $0^{\circ}\text{C}$  for a period of 6 months. At the end of this time, after the strawberries were thawed, a technician rated each strawberry's discoloration from 1 to 10, with a low score indicating little discoloration.

The dotplots below show the distributions of discoloration rating for the control and treatment groups.



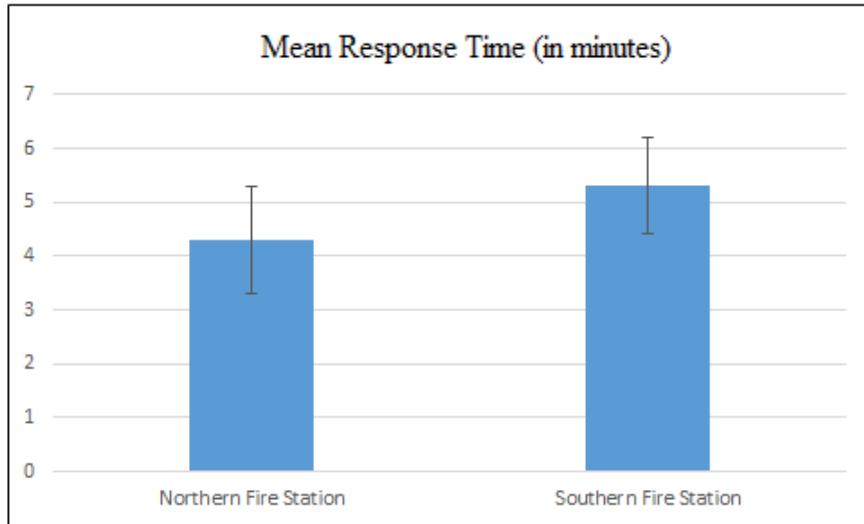
- The standard deviation of ratings for the control group is 2.141. Explain how this value summarizes variability in the control group.
- Based on the dotplots, comment on the effectiveness of the preservative in lowering the amount of discoloration in strawberries. (No calculations are necessary.)
- Researchers at the university decided to calculate a 95 percent confidence interval for the difference in mean discoloration rating between strawberries that were not treated with preservative and those that were treated with preservative. The confidence interval they obtained was (0.16, 2.72). Assume that the conditions necessary for the  $t$ -confidence interval are met.

Based on the confidence interval, comment on whether there would be a difference in the population mean discoloration ratings for the treated and untreated strawberries.

*Question 4 from 2009 Exam*

One of the two fire stations in a certain town responds to calls in the northern half of the town, and the other fire station responds to calls in the southern half of the town. One of the town council members believes that the two fire stations have different mean response times. Response time is measured by the difference between the time an emergency call comes into the fire station and the time the first fire truck arrives at the scene of the fire.

Data were collected to investigate whether the council member's belief is correct. A random sample of 50 calls selected from the northern fire station had a mean response time of 4.3 minutes with a standard deviation of 3.7 minutes. A random sample of 50 calls selected from the southern fire station had a mean response time of 5.3 minutes with a standard deviation of 3.2 minutes.



- Construct and interpret a 95 percent confidence interval for the difference in mean response times between the two fire stations.
- Does the confidence interval in part (a) support the council member's belief that the two fire stations have different mean response times? Explain.

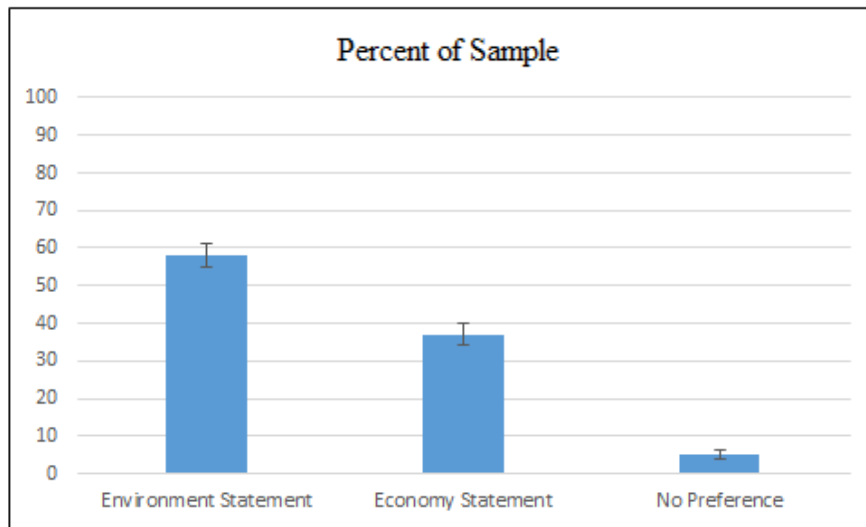
Question 5 from 2016 Exam

A polling agency showed the following two statements to a random sample of 1,048 adults in the United States.

Environment statement: Protection of the environment should be given priority over economic growth.  
Economy statement: Economic growth should be given priority over protection of the environment.

The order in which the statements were shown was randomly selected for each person in the sample. After reading the statements, each person was asked to choose the statement that was most consistent with his or her opinion. The results are shown in the table.

	Environment Statement	Economy Statement	No Preference
Percent of sample	58%	37%	5%



- (a) Assume the conditions for inference have been met. Construct and interpret a 95 percent confidence interval for the proportion of all adults in the United States who would have chosen the economy statement.
- (b) One of the conditions for inference that was met is that the number who chose the economy statement and the number who did not choose the economy statement are both greater than 10. Explain why it is necessary to satisfy that condition.
- (c) A suggestion was made to use a two-sample  $z$ -interval for a difference between proportions to investigate whether the difference in proportions between adults in the United States who would have chosen the environment statement and adults in the United States who would have chosen the economy statement is statistically significant. Is the two-sample  $z$ -interval for a difference between proportions an appropriate procedure to investigate the difference? Justify your answer.

## Appendix A4

### Confidence Interval Questions with Alternative Visual Representation

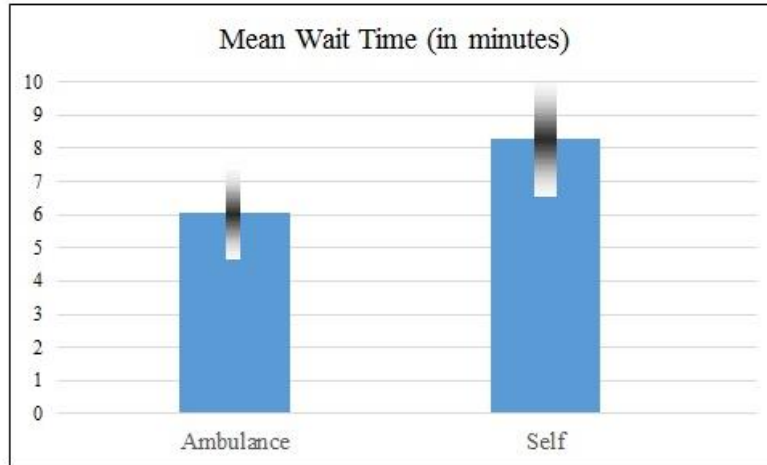
#### Question 4 from 2006 Exam

Patients with heart-attack symptoms arrive at an emergency room either by ambulance or self-transportation provided by themselves, family, or friends. When a patient arrives at the emergency room, the time of arrival is recorded. The time when the patient's diagnostic treatment begins is also recorded.

An administrator of a large hospital wanted to determine whether the mean wait time (time between arrival and diagnostic treatment) for patients with heart-attack symptoms differs according to the mode of transportation. A random sample of 150 patients with heart-attack symptoms who had reported to the emergency room was selected. For each patient, the mode of transportation and wait time were recorded. Summary statistics for each mode of transportation are shown below.

#### Mode of Transportation

Ambulance, Sample Size = 77    Mean Wait Time (in minutes) = 6.04    Standard Deviation of Wait Times (in minutes) = 4.30  
Self, Sample Size = 73    Mean Wait Time (in minutes) = 8.30    Standard Deviation of Wait Times (in minutes) = 5.16



- Use a 99 percent confidence interval to estimate the difference between the mean wait times for ambulance-transported patients and self-transported patients at this emergency room.
- Based only on this confidence interval, do you think the difference in the mean wait times is statistically significant? Justify your answer.

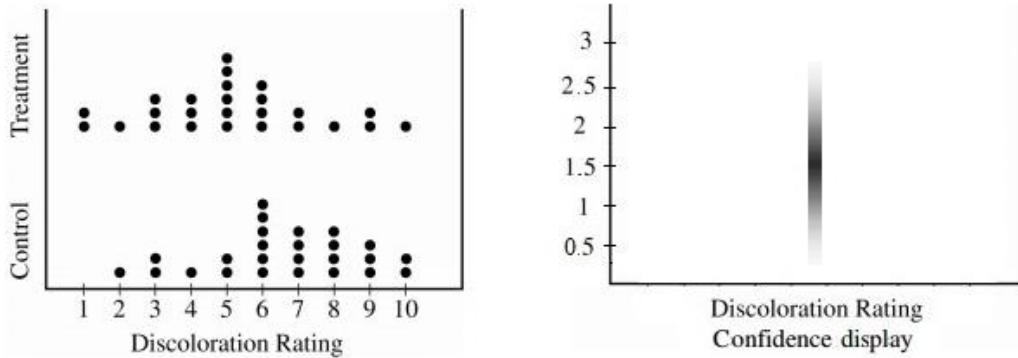
Question 1 from 2007 Exam

**Directions:** Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

- The department of agriculture at a university was interested in determining whether a preservative was effective in reducing discoloration in frozen strawberries. A sample of 50 ripe strawberries was prepared for freezing. Then the sample was randomly divided into two groups of 25 strawberries each. Each strawberry was placed into a small plastic bag.

The 25 bags in the control group were sealed. The preservative was added to the 25 bags containing strawberries in the treatment group, and then those bags were sealed. All bags were stored at 0°C for a period of 6 months. At the end of this time, after the strawberries were thawed, a technician rated each strawberry's discoloration from 1 to 10, with a low score indicating little discoloration.

The dotplots below show the distributions of discoloration rating for the control and treatment groups.



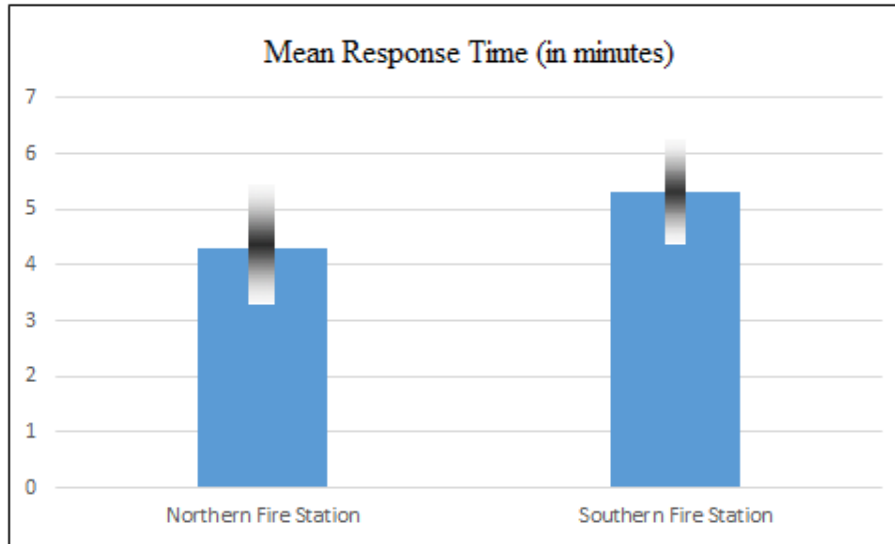
- The standard deviation of ratings for the control group is 2.141. Explain how this value summarizes variability in the control group.
- Based on the dotplots, comment on the effectiveness of the preservative in lowering the amount of discoloration in strawberries. (No calculations are necessary.)
- Researchers at the university decided to calculate a 95 percent confidence interval for the difference in mean discoloration rating between strawberries that were not treated with preservative and those that were treated with preservative. The confidence interval they obtained was (0.16, 2.72). Assume that the conditions necessary for the  $t$ -confidence interval are met.

Based on the confidence interval, comment on whether there would be a difference in the population mean discoloration ratings for the treated and untreated strawberries.

*Question 4 from 2009 Exam*

One of the two fire stations in a certain town responds to calls in the northern half of the town, and the other fire station responds to calls in the southern half of the town. One of the town council members believes that the two fire stations have different mean response times. Response time is measured by the difference between the time an emergency call comes into the fire station and the time the first fire truck arrives at the scene of the fire.

Data were collected to investigate whether the council member's belief is correct. A random sample of 50 calls selected from the northern fire station had a mean response time of 4.3 minutes with a standard deviation of 3.7 minutes. A random sample of 50 calls selected from the southern fire station had a mean response time of 5.3 minutes with a standard deviation of 3.2 minutes.



- Construct and interpret a 95 percent confidence interval for the difference in mean response times between the two fire stations.
- Does the confidence interval in part (a) support the council member's belief that the two fire stations have different mean response times? Explain.

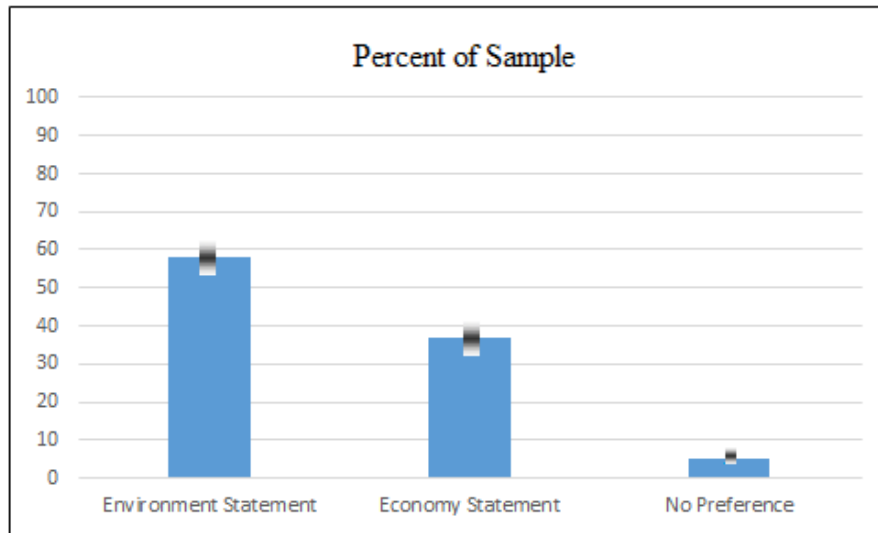
Question 5 from 2016 Exam

A polling agency showed the following two statements to a random sample of 1,048 adults in the United States.

Environment statement: Protection of the environment should be given priority over economic growth.  
Economy statement: Economic growth should be given priority over protection of the environment.

The order in which the statements were shown was randomly selected for each person in the sample. After reading the statements, each person was asked to choose the statement that was most consistent with his or her opinion. The results are shown in the table.

	Environment Statement	Economy Statement	No Preference
Percent of sample	58%	37%	5%



- Assume the conditions for inference have been met. Construct and interpret a 95 percent confidence interval for the proportion of all adults in the United States who would have chosen the economy statement.
- One of the conditions for inference that was met is that the number who chose the economy statement and the number who did not choose the economy statement are both greater than 10. Explain why it is necessary to satisfy that condition.
- A suggestion was made to use a two-sample  $z$ -interval for a difference between proportions to investigate whether the difference in proportions between adults in the United States who would have chosen the environment statement and adults in the United States who would have chosen the economy statement is statistically significant. Is the two-sample  $z$ -interval for a difference between proportions an appropriate procedure to investigate the difference? Justify your answer.

## Appendix A5

### Probability Questions with No Graphs

#### *Question 3 from 2008 Exam*

A local arcade is hosting a tournament in which contestants play an arcade game with possible scores ranging from 0 to 20. The arcade has set up multiple game tables so that all contestants can play the game at the same time; thus contestant scores are independent. Each contestant's score will be recorded as he or she finishes, and the contestant with the highest score is the winner.

After practicing the game many times, Josephine, one of the contestants, has established the probability distribution of her scores, shown in the table below.

Josephine's Distribution				
Score	16	17	18	19
Probability	0.10	0.30	0.40	0.20

Crystal, another contestant, has also practiced many times. The probability distribution for her scores is shown in the table below.

Crystal's Distribution			
Score	17	18	19
Probability	0.45	0.40	0.15

- Calculate the expected score for each player.
- Suppose that Josephine scores 16 and Crystal scores 17. The difference (Josephine minus Crystal) of their scores is  $-1$ . List all combinations of possible scores for Josephine and Crystal that will produce a difference (Josephine minus Crystal) of  $-1$ , and calculate the probability for each combination.
- Find the probability that the difference (Josephine minus Crystal) in their scores is  $-1$ .
- The table below lists all the possible differences in the scores between Josephine and Crystal and some associated probabilities.

Distribution (Josephine minus Crystal)						
Difference	-3	-2	-1	0	1	2
Probability	0.015			0.325	0.260	0.090

Complete the table and calculate the probability that Crystal's score will be higher than Josephine's score.

*Question 3 from 2015 Exam*

A shopping mall has three automated teller machines (ATMs). Because the machines receive heavy use, they sometimes stop working and need to be repaired. Let the random variable  $X$  represent the number of ATMs that are working when the mall opens on a randomly selected day. The table shows the probability distribution of  $X$ .

Number of ATMs working when the mall opens	0	1	2	3
Probability	0.15	0.21	0.40	0.24

- (a) What is the probability that at least one ATM is working when the mall opens?
- (b) What is the expected value of the number of ATMs that are working when the mall opens?
- (c) What is the probability that all three ATMs are working when the mall opens, given that at least one ATM is working?
- (d) Given that at least one ATM is working when the mall opens, would the expected value of the number of ATMs that are working be less than, equal to, or greater than the expected value from part (b) ? Explain.

## Appendix A6

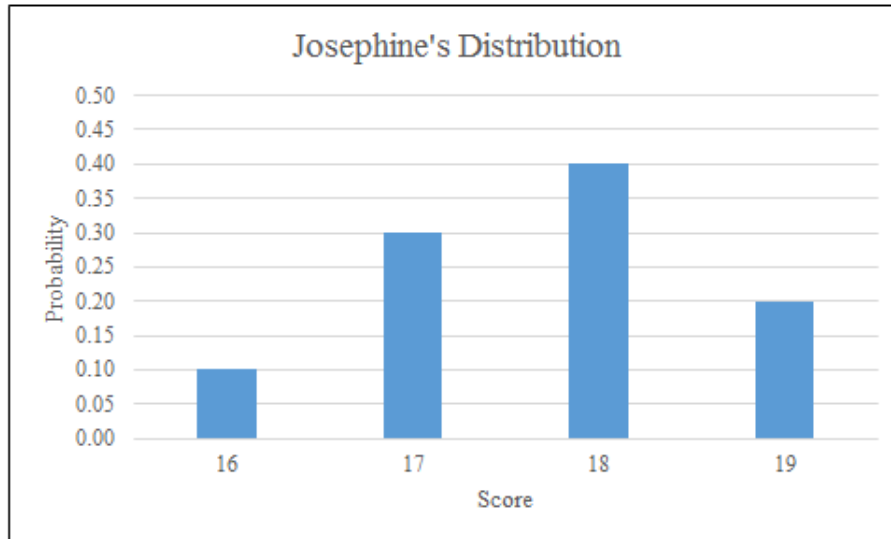
### Probability Questions with Graphs

#### *Question 3 from 2008 Exam*

A local arcade is hosting a tournament in which contestants play an arcade game with possible scores ranging from 0 to 20. The arcade has set up multiple game tables so that all contestants can play the game at the same time; thus contestant scores are independent. Each contestant's score will be recorded as he or she finishes, and the contestant with the highest score is the winner.

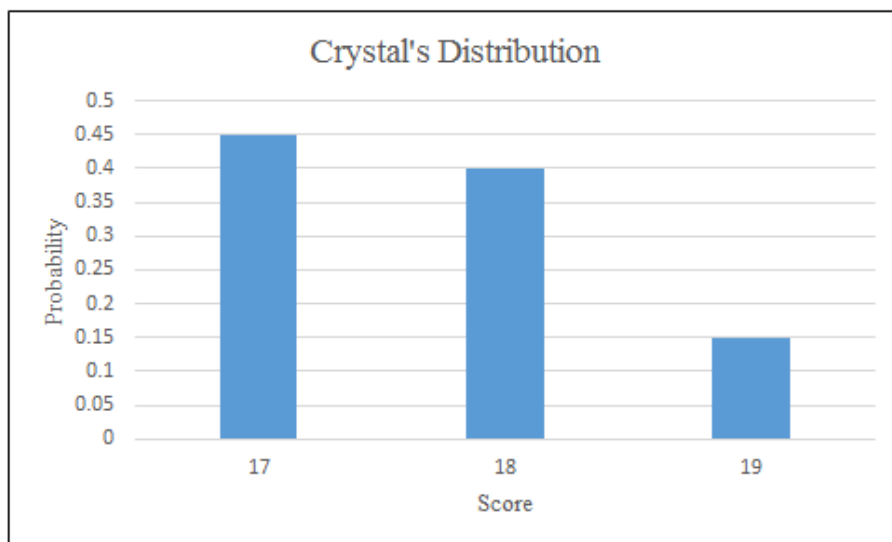
After practicing the game many times, Josephine, one of the contestants, has established the probability distribution of her scores, shown in the table below.

Josephine's Distribution				
Score	16	17	18	19
Probability	0.10	0.30	0.40	0.20



Crystal, another contestant, has also practiced many times. The probability distribution for her scores is shown in the table below.

Crystal's Distribution			
Score	17	18	19
Probability	0.45	0.40	0.15



- Calculate the expected score for each player.
- Suppose that Josephine scores 16 and Crystal scores 17. The difference (Josephine minus Crystal) of their scores is  $-1$ . List all combinations of possible scores for Josephine and Crystal that will produce a difference (Josephine minus Crystal) of  $-1$ , and calculate the probability for each combination.
- Find the probability that the difference (Josephine minus Crystal) in their scores is  $-1$ .
- The table below lists all the possible differences in the scores between Josephine and Crystal and some associated probabilities.

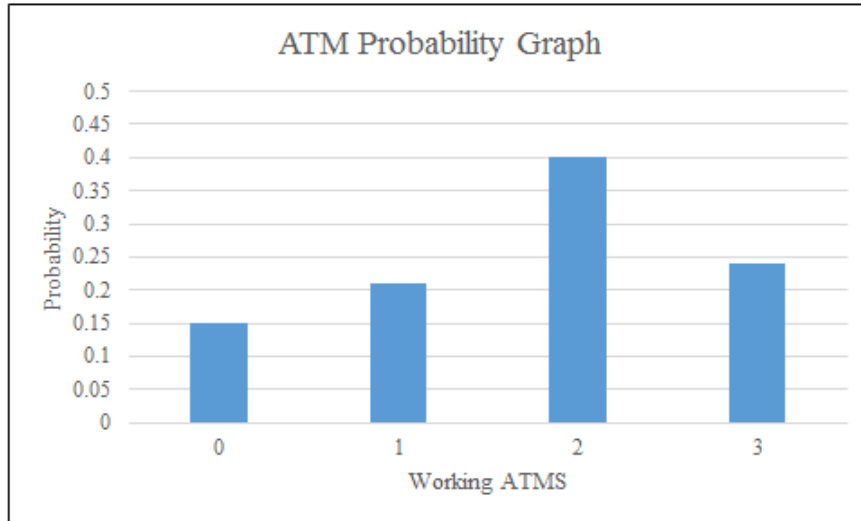
Distribution (Josephine minus Crystal)						
Difference	-3	-2	-1	0	1	2
Probability	0.015			0.325	0.260	0.090

Complete the table and calculate the probability that Crystal's score will be higher than Josephine's score.

Question 3 from 2015 Exam

A shopping mall has three automated teller machines (ATMs). Because the machines receive heavy use, they sometimes stop working and need to be repaired. Let the random variable  $X$  represent the number of ATMs that are working when the mall opens on a randomly selected day. The table shows the probability distribution of  $X$ .

Number of ATMs working when the mall opens	0	1	2	3
Probability	0.15	0.21	0.40	0.24



- (a) What is the probability that at least one ATM is working when the mall opens?
- (b) What is the expected value of the number of ATMs that are working when the mall opens?
- (c) What is the probability that all three ATMs are working when the mall opens, given that at least one ATM is working?
- (d) Given that at least one ATM is working when the mall opens, would the expected value of the number of ATMs that are working be less than, equal to, or greater than the expected value from part (b) ? Explain.

## Appendix B:

### Appendix B1: Confidence Interval Training

#### *What are Confidence Intervals?*

Confidence intervals are measures of the amount of uncertainty associated with a sample estimate of a particular population parameter. Assuming enough participants (i.e. statistical power), they can be used to infer the likely behavior of a given population.

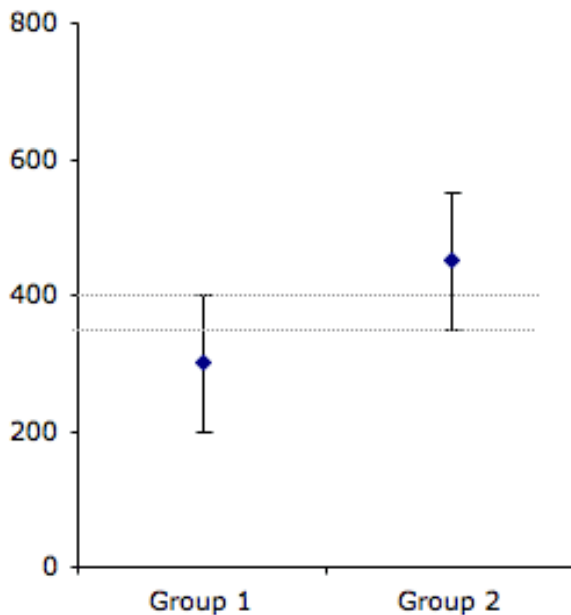
Visually, confidence intervals can be represented a variety of ways, but the most common way is using an error bar (see below):



Error bars are used to visually encapsulate the uncertainty around a data point. The most common confidence interval displayed with error bar is the 0.95 or 95% confidence interval (other values such as 0.99/99% and 0.90/90% are also used, but less frequently). The reason for this is largely due to convention and tradition, as most branches of science adhere to  $p = 0.05$  cutoff for statistical significance, which corresponds with that level of confidence interval.

#### *How they work in practice, an example:*

Let's say we measure the reaction times of 50 women in order to make generalizations about reaction times of all the women in the world. The true mean reaction time for all women is unknowable, but when we speak of a 95 percent confidence interval around our mean for the 50 women we happened to test, we are saying that if we repeatedly studied a different random sample of 50 women, 95 percent of the time, the true mean for all women will fall within the confidence interval.



**Figure 1:** Mean reaction time (ms) and 95% confidence intervals for Group 1 (n=36) and Group 2 (n=34).

Now suppose we want to know if men’s reaction times are different from women’s reaction times. We can study 50 men, compute the 95 percent confidence interval, and compare the two means and their respective confidence intervals, perhaps in a graph that looks very similar to Figure 1 above. If Group 1 is women and Group 2 is men, then the graph is saying that there’s a 95 percent chance that the true mean for all women falls within the confidence interval for Group 1, and a 95 percent chance that the true mean for all men falls within the confidence interval for Group 2.

The question is, how close can the confidence intervals be to each other and still show a significant difference? The answer is that they can overlap by as much as 25 percent of their total length and still show a significant difference between the means for each group (as shown in Figure 1). Any more overlap and the results will not be significant.

*Gradient Representations:*

Gradient representations were created as another way to represent uncertainty around a data point. On the most basic level, they operate in the same way that error bars do, but provide additional information. The darker the shades around the mean represent areas of greater certainty and the increasingly lighter shades represent a decreasing levels of certainty as you move away from the mean. In addition, gradient representations will be larger if the uncertainty around a data point is larger.

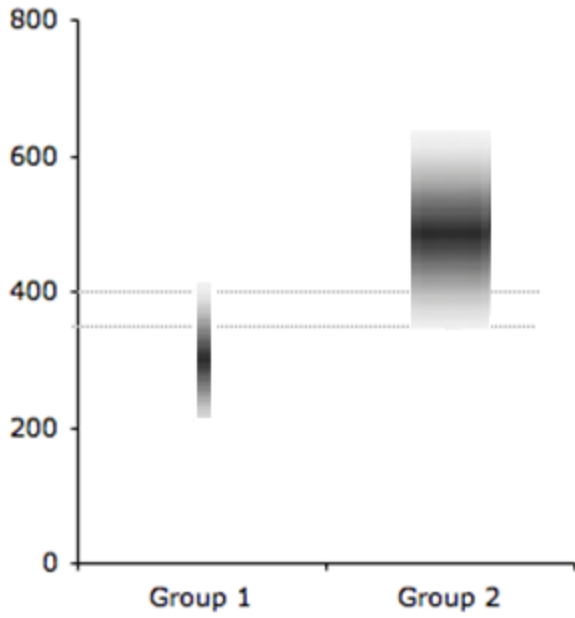


Figure 2

For example, take the scenario demonstrated in Figure 2. Size alone lets you know that Group 1 has much more certain data than Group 2. This is also communicated by decreased clarity in Group 2's display compared to Group 1. The overlap rule is identical to the error bar representation in that up to 25% overlap can occur while still indicating statistical significance (as it does in Figure 2).

## Appendix B2. Probability Training:

(Taken from <http://stattrek.com/probability/what-is-probability.aspx?Tutorial=AP>)

(\* Note: This tutorial was chosen due to the concise nature in which it covers the aspects of probability most relevant to the AP exam, its widespread usage, and most importantly, the accuracy of their tutorials.)

The probability of an event refers to the likelihood that the event will occur.

### How to Interpret Probability

Mathematically, the probability that an event will occur is expressed as a number between 0 and

1. Notationally, the probability of event A is represented by  $P(A)$ .

- If  $P(A)$  equals zero, event A will almost definitely not occur.
- If  $P(A)$  is close to zero, there is only a small chance that event A will occur.
- If  $P(A)$  equals 0.5, there is a 50-50 chance that event A will occur.
- If  $P(A)$  is close to one, there is a strong chance that event A will occur.
- If  $P(A)$  equals one, event A will almost definitely occur.
- 

In a statistical experiment, the sum of probabilities for all possible outcomes is equal to one. This means, for example, that if an experiment can have three possible outcomes (A, B, and C), then  $P(A) + P(B) + P(C) = 1$ .

### How to Compute Probability:

#### Equally Likely Outcomes

Sometimes, a statistical experiment can have  $n$  possible outcomes, each of which is equally likely. Suppose a subset of  $r$  outcomes are classified as "successful" outcomes.

The probability that the experiment results in a successful outcome (S) is:

$$P(S) = (\text{Number of successful outcomes}) / (\text{Total number of equally likely outcomes}) = r / n$$

Consider the following experiment:

An urn has 10 marbles. Two marbles are red, three are green, and five are blue. If an experimenter randomly selects 1 marble from the urn, what is the probability that it will be green?

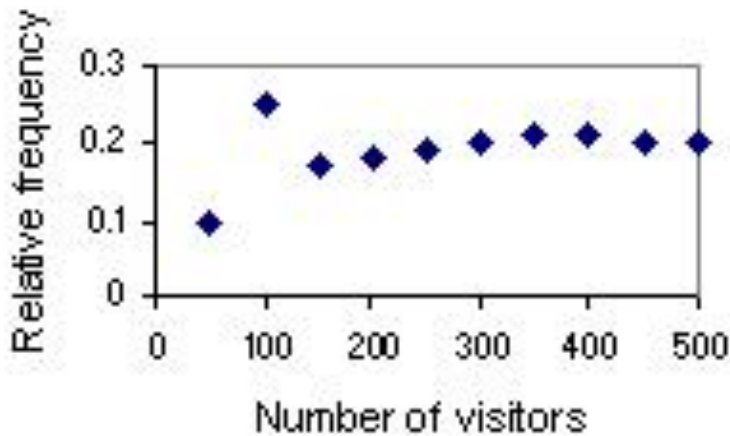
In this experiment, there are 10 equally likely outcomes, three of which are green marbles. Therefore, the probability of choosing a green marble is  $3/10$  or 0.30.

### How to Compute Probability: Law of Large Numbers

One can also think about the probability of an event in terms of its *long-run* relative frequency. The relative frequency of an event is the number of times an event occurs, divided by the total number of trials.

$$P(A) = (\text{Frequency of Event A}) / (\text{Number of Trials})$$

For example, a merchant notices one day that 5 out of 50 visitors to her store make a purchase. The next day, 20 out of 50 visitors make a purchase. The two relative frequencies (5/50 or 0.10 and 20/50 or 0.40) differ. However, summing results over many visitors, she might find that the probability that a visitor makes a purchase gets closer and closer 0.20.



The scatterplot shows the relative frequency as the number of trials (in this case, the number of visitors) increases. Over many trials, the relative frequency converges toward a stable value (0.20), which can be interpreted as the probability that a visitor to the store will make a purchase. The idea that the relative frequency of an event will converge on the probability of the event, as the number of trials increases, is called the law of large numbers.

Appendix C.

Appendix C1.

*MathML Code*

```
<!DOCTYPE html>
<html>
<body>

<h1>Formulas</h1>

<h2>Descriptive Statistics</h2>

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <mover>
      <mi>x</mi>
      <mo>¯ </mo>
    </mover>
    <mo>=</mo>
    <mfrac>
      <mrow>
        <mo>∑</mo>
        <msub>
          <mi>x</mi>
          <mi>i</mi>
        </msub>
      </mrow>
      <mi>n</mi>
    </mfrac>
  </mstyle>
</math>
```

```
<br><br>
```

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <msub>
      <mi>s</mi>
      <mi>x</mi>
    </msub>
    <mo>=</mo>
    <msqrt>
      <mrow>
        <mfrac>
```

```

<mn>1</mn>
<mrow>
  <mi>n</mi>
  <mo>-</mo>
  <mn>1</mn>
</mrow>
</mfrac>
<mo>#x2211;</mo>
<msup>
  <mrow>
    <mo>(</mo>
    <msub>
      <mi>x</mi>
      <mi>i</mi>
    </msub>
    <mo>-</mo>
  </mrow>
  <mover>
    <mi>x</mi>
    <mo>#xAF;</mo>
  </mover>
  <mo>)</mo>
</msup>
<mn>2</mn>
</mrow>
</msqrt>
</mstyle>
</math>

```

```
<br><br>
```

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <msub>
      <mi>s</mi>
      <mi>p</mi>
    </msub>
    <mo>=</mo>
    <msqrt>
      <mrow>
        <mfrac>
          <mrow>
            <mrow>
              <mo>(</mo>
              <msub>
                <mi>n</mi>

```

```

    <mn>1</mn>
  </msub>
  <mo>-</mo>
  <mn>1</mn>
  <mo>)</mo>
</mrow>
<mrow>
  <msubsup>
    <mi>s</mi>
    <mn>1</mn>
    <mn>2</mn>
  </msubsup>
</mrow>
<mo>+</mo>
<mrow>
  <mo>(</mo>
  <msub>
    <mi>n</mi>
    <mn>2</mn>
  </msub>
  <mo>-</mo>
  <mn>1</mn>
  <mo>)</mo>
</mrow>
<mrow>
  <msubsup>
    <mi>s</mi>
    <mn>2</mn>
    <mn>2</mn>
  </msubsup>
</mrow>
</mrow>
<mrow>
  <mrow>
    <mo>(</mo>
    <msub>
      <mi>n</mi>
      <mn>1</mn>
    </msub>
    <mo>-</mo>
    <mn>1</mn>
    <mo>)</mo>
  </mrow>
  <mo>-</mo>
</mrow>
  <mo>(</mo>

```

```

    <msub>
      <mi>n</mi>
      <mn>2</mn>
    </msub>
    <mo>-</mo>
    <mn>1</mn>
    <mo>)</mo>
  </mrow>
</mrow>
</mfrac>
</mrow>
</msqrt>
</mstyle>
</math>

```

```
<br><br>
```

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <mover>
      <mi>y</mi>
      <mo>^</mo>
    </mover>
    <mo>=</mo>
    <msub>
      <mi>b</mi>
      <mn>0</mn>
    </msub>
    <mo>+</mo>
    <msub>
      <mi>b</mi>
      <mn>1</mn>
    </msub>
    <mi>x</mi>
  </mstyle>
</math>

```

```
<br><br>
```

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <msub>
      <mi>b</mi>
      <mn>1</mn>
    </msub>
    <mo>=</mo>

```

```

<mfrac>
<mrow>
<mo>#x2211;</mo>
<mrow>
<mo>(</mo>
<msub>
<mi>x</mi>
<mi>i</mi>
</msub>
<mo>-</mo>
<mover>
<mi>x</mi>
<mo>#xAF;</mo>
</mover>
<mo>)</mo>
</mrow>
<mrow>
<mo>(</mo>
<msub>
<mi>y</mi>
<mi>i</mi>
</msub>
<mo>-</mo>
<mover>
<mi>y</mi>
<mo>#xAF;</mo>
</mover>
<mo>)</mo>
</mrow>
</mrow>
<mrow>
<mo>#x2211;</mo>
<msup>
<mrow>
<mo>(</mo>
<msub>
<mi>x</mi>
<mi>i</mi>
</msub>
<mo>-</mo>
<mover>
<mi>x</mi>
<mo>#xAF;</mo>
</mover>
<mo>)</mo>
</mrow>

```

```
<mn>2</mn>
</msup>
</mrow>
</mfrac>
</mstyle>
</math>
```

```
<br><br>
```

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <msub>
      <mi>b</mi>
      <mn>0</mn>
    </msub>
    <mo>=</mo>
    <mover>
      <mi>y</mi>
      <mo>#xAF;</mo>
    </mover>
    <mo>-</mo>
    <msub>
      <mi>b</mi>
      <mn>1</mn>
    </msub>
    <mover>
      <mi>x</mi>
      <mo>#xAF;</mo>
    </mover>
  </mstyle>
</math>
```

```
<br><br>
```

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <mi>r</mi>
    <mo>=</mo>
    <mfrac>
      <mn>1</mn>
      <mrow>
        <mi>n</mi>
        <mo>-</mo>
        <mn>1</mn>
      </mrow>
    </mfrac>
```

```

<mo>&#x2211;</mo>
<mrow>
  <mo></mo>
  <mfrac>
    <mrow>
      <msub>
        <mi>x</mi>
        <mi>i</mi>
      </msub>
      <mo>-</mo>
      <mover>
        <mi>x</mi>
        <mo>&#xAF;</mo>
      </mover>
    </mrow>
    <msub>
      <mi>s</mi>
      <mi>x</mi>
    </msub>
  </mfrac>
  <mo></mo>
</mrow>
<mrow>
  <mo></mo>
  <mfrac>
    <mrow>
      <msub>
        <mi>y</mi>
        <mi>i</mi>
      </msub>
      <mo>-</mo>
      <mover>
        <mi>y</mi>
        <mo>&#xAF;</mo>
      </mover>
    </mrow>
    <msub>
      <mi>s</mi>
      <mi>y</mi>
    </msub>
  </mfrac>
  <mo></mo>
</mrow>
</mstyle>
</math>

```

<br><br>

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <msub>
      <mi>b</mi>
      <mn>1</mn>
    </msub>
    <mo>=</mo>
    <mi>r</mi>
    <mrow>
      <mo>(</mo>
      <mfrac>
        <msub>
          <mi>s</mi>
          <mi>y</mi>
        </msub>
        <msub>
          <mi>s</mi>
          <mi>x</mi>
        </msub>
      </mfrac>
      <mo>)</mo>
    </mrow>
  </mstyle>
</math>
```

<br><br>

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <msub>
      <mi>s</mi>
      <mrow>
        <msub>
          <mi>b</mi>
          <mn>1</mn>
        </msub>
      </mrow>
    </msub>
    <mo>=</mo>
    <mfrac>
      <mrow>
        <mfrac>
          <msqrt>
            <mrow>
```

```

<mo>#x2211;</mo>
<msup>
  <mrow>
    <mo>(</mo>
    <msub>
      <mi>y</mi>
      <mi>i</mi>
    </msub>
    <mo>-</mo>
    <msub>
      <mover>
        <mi>y</mi>
        <mo>^</mo>
      </mover>
      <mi>i</mi>
    </msub>
    <mo>)</mo>
  </mrow>
  <mn>2</mn>
</msup>
</mrow>
</msqrt>
<mrow>
  <mi>n</mi>
  <mo>-</mo>
  <mn>2</mn>
</mrow>
</mfrac>
</mrow>
<mrow>
  <msqrt>
    <mrow>
      <mo>#x2211;</mo>
      <msup>
        <mrow>
          <mo>(</mo>
          <msub>
            <mi>x</mi>
            <mi>i</mi>
          </msub>
          <mo>-</mo>
          <mover>
            <mi>x</mi>
            <mo>#xAF;</mo>
          </mover>
          <mo>)</mo>
        </mrow>
      </msup>
    </mrow>
  </msqrt>
</mrow>

```

```
</mrow>
<mn>2</mn>
</msup>
</mrow>
</msqrt>
</mrow>
</mfrac>
</mstyle>
</math>
```

```
<br><br>
```

```
<h3>Probability</h3>
```

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
<mstyle displaystyle="true">
<mi>P</mi>
<mrow>
<mo>(</mo>
<mi>a</mi>
<mo>&#x222A;</mo>
<mi>b</mi>
<mo>)</mo>
</mrow>
<mo>=</mo>
<mi>P</mi>
<mrow>
<mo>(</mo>
<mi>A</mi>
<mo>)</mo>
</mrow>
<mo>+</mo>
<mi>P</mi>
<mrow>
<mo>(</mo>
<mi>B</mi>
<mo>)</mo>
</mrow>
<mo>-</mo>
<mrow>
<mo>(</mo>
<mi>A</mi>
<mo>&#x2229;</mo>
<mi>B</mi>
<mo>)</mo>
</mrow>
```

</mstyle>  
</math>

<br><br>

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <mi>P</mi>
    <mrow>
      <mo>(</mo>
      <mi>A</mi>
      <mrow>
        <mo>&#x2223;</mo>
      </mrow>
      <mi>B</mi>
      <mo>)</mo>
    </mrow>
    <mo>=</mo>
    <mfrac>
      <mrow>
        <mi>P</mi>
        <mrow>
          <mo>(</mo>
          <mi>A</mi>
          <mo>&#x2229;</mo>
          <mi>B</mi>
          <mo>)</mo>
        </mrow>
      </mrow>
      <mrow>
        <mi>P</mi>
        <mrow>
          <mo>(</mo>
          <mi>B</mi>
          <mo>)</mo>
        </mrow>
      </mrow>
    </mfrac>
  </mstyle>
</math>
```

<br><br>

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <mi>E</mi>
  </mstyle>
</math>
```

```

<mrow>
  <mo>(</mo>
  <mi>X</mi>
  <mo>)</mo>
</mrow>
<mo>=</mo>
<msub>
  <mi>&#x3BC;</mi>
  <mi>x</mi>
</msub>
<mo>=</mo>
<mo>&#x2211;</mo>
<msub>
  <mi>x</mi>
  <mi>i</mi>
</msub>
<msub>
  <mi>p</mi>
  <mi>i</mi>
</msub>
</mstyle>
</math>

```

```
<br><br>
```

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <mi>V</mi>
    <mi>a</mi>
    <mi>r</mi>
    <mrow>
      <mo>(</mo>
      <mi>X</mi>
      <mo>)</mo>
    </mrow>
    <mo>=</mo>
    <mrow>
      <msubsup>
        <mi>&#x3C3;</mi>
        <mi>x</mi>
        <mn>2</mn>
      </msubsup>
    </mrow>
    <mo>=</mo>
    <mo>&#x2211;</mo>
    <msup>

```

```

<mrow>
  <mo>(</mo>
  <msub>
    <mi>x</mi>
    <mi>i</mi>
  </msub>
  <mo>-</mo>
  <msub>
    <mi>#x3BC;</mi>
    <mi>x</mi>
  </msub>
  <mo>)</mo>
</mrow>
<mn>2</mn>
</msup>
<msub>
  <mi>p</mi>
  <mi>i</mi>
</msub>
</mstyle>
</math>

```

```
<br><br>
```

<h4>If X has a binomial distribution with parameters n and p, then:</h4>

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <mi>P</mi>
    <mrow>
      <mo>(</mo>
      <mi>X</mi>
      <mo>=</mo>
      <mi>k</mi>
      <mo>)</mo>
    </mrow>
    <mo>=</mo>
    <mrow>
      <mo>(</mo>
      <mtable>
        <mtr>
          <mtd>
            <mi>n</mi>
          </mtd>
        </mtr>
      </mtable>
    </mrow>
  </mstyle>

```

```

    <mtd>
      <mi>k</mi>
    </mtd>
  </mtr>
</mtable>
<mo>)</mo>
</mrow>
<msup>
  <mi>p</mi>
  <mi>k</mi>
</msup>
<msup>
  <mrow>
    <mo>(</mo>
    <mn>1</mn>
    <mo>-</mo>
    <mi>p</mi>
    <mo>)</mo>
  </mrow>
  <mrow>
    <mi>n</mi>
    <mo>-</mo>
    <mi>k</mi>
  </mrow>
</msup>
</mstyle>
</math>

```

```
<br><br>
```

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <msub>
      <mi>&#x3BC;</mi>
      <mi>x</mi>
    </msub>
    <mo>=</mo>
    <mi>n</mi>
    <mi>p</mi>
  </mstyle>
</math>

```

```
<br><br>
```

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">

```

```

<msub>
  <mi>#x3C3;</mi>
  <mi>x</mi>
</msub>
<mo>=</mo>
<msqrt>
  <mrow>
    <mi>n</mi>
    <mi>p</mi>
  </mrow>
  <mo>(</mo>
  <mn>1</mn>
  <mo>-</mo>
  <mi>p</mi>
  <mo>)</mo>
</mrow>
</msqrt>
</mstyle>
</math>

```

```
<br><br>
```

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <msub>
      <mi>#x3BC;</mi>
      <mover>
        <mi>p</mi>
        <mo>^</mo>
      </mover>
    </msub>
    <mo>=</mo>
    <mi>p</mi>
  </mstyle>
</math>

```

```
<br><br>
```

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <msub>
      <mi>#x3C3;</mi>
      <mover>
        <mi>p</mi>
        <mo>^</mo>

```

```

    </mover>
  </msub>
  <mo>=</mo>
  <msqrt>
    <mrow>
      <mfrac>
        <mrow>
          <mi>p</mi>
          <mrow>
            <mo>(</mo>
            <mn>1</mn>
            <mo>-</mo>
            <mi>p</mi>
            <mo>)</mo>
          </mrow>
        </mrow>
      </mfrac>
      <mi>n</mi>
    </mrow>
  </msqrt>
</mstyle>
</math>

```

```
<br><br>
```

<h5><font size="3">If  $\bar{x}$ ; is the mean of a random sample of size  $n$  from an infinite population with mean  $\mu$ ; and standard deviation  $\sigma$ ; then:</font></h5>

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <msub>
      <mi>&#x3BC;</mi>
      <mover>
        <mi>x</mi>
        <mo>&#xAF;</mo>
      </mover>
    </msub>
    <mo>=</mo>
    <mi>&#x3BC;</mi>
  </mstyle>
</math>

```

```
<br><br>
```

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">

```

```

<msub>
  <mi>#x3C3;</mi>
  <mover>
    <mi>x</mi>
    <mo>#xAF;</mo>
  </mover>
</msub>
<mo>=</mo>
<mfrac>
  <mi>#x3C3;</mi>
  <msqrt>
    <mi>n</mi>
  </msqrt>
</mfrac>
</mstyle>
</math>

```

<h6><font size="3">Inferential Statistics</font></h6>

<body>Standardized test statistic: (statistics-parameter)&#247;(standard deviation of statistic)  
</body>

<br><br>

<body>Confidence interval: statistic &#177; (critical value) &times; (standard deviation of statistic) </body>

<br><br>

<h7>Single-Sample:</h7>

<br><br>

<body>Sample Mean Standard Deviation</body>

<br>

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <mfrac>
      <mi>#x3C3;</mi>
      <msqrt>
        <mi>n</mi>
      </msqrt>
    </mfrac>
  </mstyle>

```

</math>

<br>

<body>Sample Proportion Standard Deviation</body>

<br><br>

<math xmlns="http://www.w3.org/1998/Math/MathML">

<mstyle displaystyle="true">

<msqrt>

<mrow>

<mfrac>

<mrow>

<mi>p</mi>

<mrow>

<mo>(</mo>

<mi>p</mi>

<mo>-</mo>

<mn>1</mn>

<mo>)</mo>

</mrow>

</mrow>

<mi>n</mi>

</mfrac>

</mrow>

</msqrt>

</mstyle>

</math>

<br><br>

<h2>Two-Sample:</h2>

<br><br>

<body>Difference of Sample Means Standard Deviation</body>

<br><br>

<math xmlns="http://www.w3.org/1998/Math/MathML">

<mstyle displaystyle="true">

<msqrt>

<mrow>

<mfrac>

<mrow>

<msubsup>

```

    <mi>∑;</mi>
    <mn>1</mn>
    <mn>2</mn>
  </msubsup>
</mrow>
<msub>
  <mi>n</mi>
  <mn>1</mn>
</msub>
</mfrac>
<mo>+</mo>
<mfrac>
  <mrow>
    <msubsup>
      <mi>∑;</mi>
      <mn>2</mn>
      <mn>2</mn>
    </msubsup>
  </mrow>
  <msub>
    <mi>n</mi>
    <mn>2</mn>
  </msub>
</mfrac>
</mrow>
</msqrt>
</mstyle>
</math>

```

```
<br>
```

```
<body> If equal ∑; use the following formula instead </body>
```

```
<br>
```

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <mi>∑;</mi>
    <msqrt>
      <mrow>
        <mfrac>
          <mn>1</mn>
          <msub>
            <mi>n</mi>
            <mn>1</mn>
          </msub>

```

```

</mfrac>
<mo>+</mo>
<mfrac>
  <mn>1</mn>
  <msub>
    <mi>n</mi>
    <mn>2</mn>
  </msub>
</mfrac>
</mrow>
</msqrt>
</mstyle>
</math>

```

```
<br><br>
```

```
<body>Difference of Sample Proportions Standard Deviation</body>
```

```
<br>
```

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <msqrt>
      <mrow>
        <mfrac>
          <mrow>
            <msub>
              <mi>p</mi>
              <mn>1</mn>
            </msub>
            <mrow>
              <mo>(</mo>
              <mn>1</mn>
              <mo>-</mo>
              <msub>
                <mi>p</mi>
                <mn>1</mn>
              </msub>
              <mo>)</mo>
            </mrow>
          </mrow>
        </mfrac>
      </msqrt>
    </mstyle>
  </math>

```

```

<mo>+</mo>
<mfrac>
  <mrow>
    <msub>
      <mi>p</mi>
      <mn>2</mn>
    </msub>
    <mrow>
      <mo>(</mo>
      <mn>1</mn>
      <mo>-</mo>
      <msub>
        <mi>p</mi>
        <mn>2</mn>
      </msub>
      <mo>)</mo>
    </mrow>
  </mrow>
</mfrac>
<msub>
  <mi>n</mi>
  <mn>2</mn>
</msub>
</math>

```

```
<br><br>
```

```
<body> If  $p_1=p_2$ ; use the following formula instead </body>
```

```
<br>
```

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <msqrt>
      <mrow>
        <mi>p</mi>
        <mrow>
          <mo>(</mo>
          <mn>1</mn>
          <mo>-</mo>
          <mi>p</mi>
          <mo>)</mo>
        </mrow>
      </mrow>
    </msqrt>
  </mstyle>
</math>

```

```

</mrow>
</msqrt>
<msqrt>
  <mrow>
    <mfrac>
      <mn>1</mn>
      <msub>
        <mi>n</mi>
        <mn>1</mn>
      </msub>
    </mfrac>
    <mo>+</mo>
    <mfrac>
      <mn>1</mn>
      <msub>
        <mi>n</mi>
        <mn>2</mn>
      </msub>
    </mfrac>
  </mrow>
</msqrt>
</mstyle>
</math>

```

```
<br>
```

```
<h9>Chi-Square test statistic</h9>
```

```

<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mstyle displaystyle="true">
    <mo>=</mo>
    <mo>&#x2211;</mo>
    <mrow>
      <mo></mo>
      <mfrac>
        <msup>
          <mrow>
            <mo></mo>
            <mi>observed</mi>
            <mo>-</mo>
            <mi>expected</mi>
            <mo></mo>
          </mrow>
          <mn>2</mn>
        </msup>
        <mi>expected</mi>
      </mfrac>
    </mrow>
  </mstyle>
</math>

```

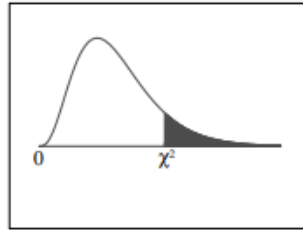
</mfrac>  
<mo>)</mo>  
</mrow>  
</mstyle>  
</math>

Appendix C2

Chi-Square Table

Note: Original file is tagged screen reader navigable pdf

Chi-Square Distribution Table



The shaded area is equal to  $\alpha$  for  $\chi^2 = \chi^2_{\alpha}$ .

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

## Z Table

*Note: Original file is tagged navigable Excel sheet*

Function: NORMDIST

Input 1: Score (Z-Score for Standard Normal Distribution)

Input 2: Mean (0 for Standard Normal Distribution)

Input 3: Standard Deviation (1 for Standard Normal Distribution)

Input 4: Cumulative (always 1)

<b>Cumulative Probability</b>	<b>Z-score</b>
2.86652E-07	-5.00
3.01896E-07	-4.99
3.17921E-07	-4.98
3.34765E-07	-4.97
3.52466E-07	-4.96
3.71067E-07	-4.95
3.90613E-07	-4.94
4.11148E-07	-4.93
4.32721E-07	-4.92
4.55382E-07	-4.91
4.79183E-07	-4.90
5.0418E-07	-4.89
5.30429E-07	-4.88
5.57991E-07	-4.87
5.86929E-07	-4.86
6.17307E-07	-4.85
6.49196E-07	-4.84
6.82665E-07	-4.83
7.17791E-07	-4.82
7.54651E-07	-4.81
7.93328E-07	-4.80
8.33907E-07	-4.79
8.76476E-07	-4.78
9.2113E-07	-4.77
9.67965E-07	-4.76
1.01708E-06	-4.75
1.06859E-06	-4.74
1.1226E-06	-4.73
1.17922E-06	-4.72
1.23858E-06	-4.71

1.30081E-06	-4.70
1.36603E-06	-4.69
1.43437E-06	-4.68
1.506E-06	-4.67
1.58105E-06	-4.66
1.65968E-06	-4.65
1.74205E-06	-4.64
1.82833E-06	-4.63
1.9187E-06	-4.62
2.01334E-06	-4.61
2.11245E-06	-4.60
2.21623E-06	-4.59
2.32488E-06	-4.58
2.43862E-06	-4.57
2.55768E-06	-4.56
2.6823E-06	-4.55
2.81271E-06	-4.54
2.94918E-06	-4.53
3.09198E-06	-4.52
3.24138E-06	-4.51
3.39767E-06	-4.50
3.56116E-06	-4.49
3.73215E-06	-4.48
3.91098E-06	-4.47
4.09798E-06	-4.46
4.29351E-06	-4.45
4.49794E-06	-4.44
4.71165E-06	-4.43
4.93505E-06	-4.42
5.16853E-06	-4.41
5.41254E-06	-4.40
5.66753E-06	-4.39
5.93397E-06	-4.38
6.21233E-06	-4.37
6.50312E-06	-4.36
6.80688E-06	-4.35
7.12414E-06	-4.34
7.45547E-06	-4.33
7.80146E-06	-4.32
8.16273E-06	-4.31
8.53991E-06	-4.30
8.93366E-06	-4.29
9.34467E-06	-4.28

9.77365E-06	-4.27
1.02213E-05	-4.26
1.06885E-05	-4.25
1.1176E-05	-4.24
1.16846E-05	-4.23
1.22151E-05	-4.22
1.27685E-05	-4.21
1.33457E-05	-4.20
1.39477E-05	-4.19
1.45755E-05	-4.18
1.523E-05	-4.17
1.59124E-05	-4.16
1.66238E-05	-4.15
1.73653E-05	-4.14
1.81382E-05	-4.13
1.89436E-05	-4.12
1.9783E-05	-4.11
2.06575E-05	-4.10
2.15687E-05	-4.09
2.25179E-05	-4.08
2.35066E-05	-4.07
2.45364E-05	-4.06
2.56088E-05	-4.05
2.67256E-05	-4.04
2.78884E-05	-4.03
2.90991E-05	-4.02
3.03594E-05	-4.01
3.16712E-05	-4.00
3.30366E-05	-3.99
3.44576E-05	-3.98
3.59363E-05	-3.97
3.74749E-05	-3.96
3.90756E-05	-3.95
4.07408E-05	-3.94
4.24729E-05	-3.93
4.42745E-05	-3.92
4.61481E-05	-3.91
4.80963E-05	-3.90
5.01221E-05	-3.89
5.22282E-05	-3.88
5.44177E-05	-3.87
5.66935E-05	-3.86
5.90589E-05	-3.85

6.15172E-05	-3.84
6.40716E-05	-3.83
6.67258E-05	-3.82
6.94834E-05	-3.81
7.2348E-05	-3.80
7.53236E-05	-3.79
7.84142E-05	-3.78
8.16238E-05	-3.77
8.49567E-05	-3.76
8.84173E-05	-3.75
9.20101E-05	-3.74
9.57399E-05	-3.73
9.96114E-05	-3.72
0.00010363	-3.71
0.0001078	-3.70
0.000112127	-3.69
0.000116617	-3.68
0.000121275	-3.67
0.000126108	-3.66
0.00013112	-3.65
0.000136319	-3.64
0.000141711	-3.63
0.000147302	-3.62
0.000153099	-3.61
0.000159109	-3.60
0.000165339	-3.59
0.000171797	-3.58
0.000178491	-3.57
0.000185427	-3.56
0.000192616	-3.55
0.000200064	-3.54
0.00020778	-3.53
0.000215773	-3.52
0.000224053	-3.51
0.000232629	-3.50
0.00024151	-3.49
0.000250707	-3.48
0.000260229	-3.47
0.000270088	-3.46
0.000280293	-3.45
0.000290857	-3.44
0.000301791	-3.43
0.000313106	-3.42

0.000324814	-3.41
0.000336929	-3.40
0.000349463	-3.39
0.000362429	-3.38
0.000375841	-3.37
0.000389712	-3.36
0.000404058	-3.35
0.000418892	-3.34
0.00043423	-3.33
0.000450087	-3.32
0.00046648	-3.31
0.000483424	-3.30
0.000500937	-3.29
0.000519035	-3.28
0.000537737	-3.27
0.000557061	-3.26
0.000577025	-3.25
0.000597648	-3.24
0.000618951	-3.23
0.000640953	-3.22
0.000663675	-3.21
0.000687138	-3.20
0.000711364	-3.19
0.000736375	-3.18
0.000762195	-3.17
0.000788846	-3.16
0.000816352	-3.15
0.000844739	-3.14
0.000874032	-3.13
0.000904255	-3.12
0.000935437	-3.11
0.000967603	-3.10
0.001000782	-3.09
0.001035003	-3.08
0.001070294	-3.07
0.001106685	-3.06
0.001144207	-3.05
0.001182891	-3.04
0.001222769	-3.03
0.001263873	-3.02
0.001306238	-3.01
0.001349898	-3.00
0.001394887	-2.99

0.001441242	-2.98
0.001488999	-2.97
0.001538195	-2.96
0.00158887	-2.95
0.001641061	-2.94
0.00169481	-2.93
0.001750157	-2.92
0.001807144	-2.91
0.001865813	-2.90
0.001926209	-2.89
0.001988376	-2.88
0.002052359	-2.87
0.002118205	-2.86
0.002185961	-2.85
0.002255677	-2.84
0.0023274	-2.83
0.002401182	-2.82
0.002477075	-2.81
0.00255513	-2.80
0.002635402	-2.79
0.002717945	-2.78
0.002802815	-2.77
0.002890068	-2.76
0.002979763	-2.75
0.003071959	-2.74
0.003166716	-2.73
0.003264096	-2.72
0.00336416	-2.71
0.003466974	-2.70
0.003572601	-2.69
0.003681108	-2.68
0.003792562	-2.67
0.003907033	-2.66
0.004024589	-2.65
0.004145301	-2.64
0.004269243	-2.63
0.004396488	-2.62
0.004527111	-2.61
0.004661188	-2.60
0.004798797	-2.59
0.004940016	-2.58
0.005084926	-2.57
0.005233608	-2.56

0.005386146	-2.55
0.005542623	-2.54
0.005703126	-2.53
0.005867742	-2.52
0.006036558	-2.51
0.006209665	-2.50
0.006387155	-2.49
0.006569119	-2.48
0.006755653	-2.47
0.006946851	-2.46
0.007142811	-2.45
0.007343631	-2.44
0.007549411	-2.43
0.007760254	-2.42
0.00797626	-2.41
0.008197536	-2.40
0.008424186	-2.39
0.008656319	-2.38
0.008894043	-2.37
0.009137468	-2.36
0.009386706	-2.35
0.00964187	-2.34
0.009903076	-2.33
0.010170439	-2.32
0.010444077	-2.31
0.01072411	-2.30
0.011010658	-2.29
0.011303844	-2.28
0.011603792	-2.27
0.011910625	-2.26
0.012224473	-2.25
0.012545461	-2.24
0.012873721	-2.23
0.013209384	-2.22
0.013552581	-2.21
0.013903448	-2.20
0.014262118	-2.19
0.014628731	-2.18
0.015003423	-2.17
0.015386335	-2.16
0.015777607	-2.15
0.016177383	-2.14
0.016585807	-2.13

0.017003023	-2.12
0.017429178	-2.11
0.017864421	-2.10
0.0183089	-2.09
0.018762766	-2.08
0.019226172	-2.07
0.01969927	-2.06
0.020182215	-2.05
0.020675163	-2.04
0.02117827	-2.03
0.021691694	-2.02
0.022215594	-2.01
0.022750132	-2.00
0.023295468	-1.99
0.023851764	-1.98
0.024419185	-1.97
0.024997895	-1.96
0.02558806	-1.95
0.026189845	-1.94
0.026803419	-1.93
0.02742895	-1.92
0.028066607	-1.91
0.02871656	-1.90
0.02937898	-1.89
0.030054039	-1.88
0.030741909	-1.87
0.031442763	-1.86
0.032156775	-1.85
0.032884119	-1.84
0.033624969	-1.83
0.034379502	-1.82
0.035147894	-1.81
0.035930319	-1.80
0.036726956	-1.79
0.03753798	-1.78
0.03836357	-1.77
0.039203903	-1.76
0.040059157	-1.75
0.040929509	-1.74
0.041815138	-1.73
0.042716221	-1.72
0.043632937	-1.71
0.044565463	-1.70

0.045513977	-1.69
0.046478658	-1.68
0.047459682	-1.67
0.048457226	-1.66
0.049471468	-1.65
0.050502583	-1.64
0.051550748	-1.63
0.052616138	-1.62
0.053698928	-1.61
0.054799292	-1.60
0.055917403	-1.59
0.057053433	-1.58
0.058207556	-1.57
0.059379941	-1.56
0.060570758	-1.55
0.061780177	-1.54
0.063008364	-1.53
0.064255488	-1.52
0.065521712	-1.51
0.066807201	-1.50
0.068112118	-1.49
0.069436623	-1.48
0.070780877	-1.47
0.072145037	-1.46
0.07352926	-1.45
0.0749337	-1.44
0.07635851	-1.43
0.077803841	-1.42
0.079269841	-1.41
0.080756659	-1.40
0.082264439	-1.39
0.083793322	-1.38
0.085343451	-1.37
0.086914962	-1.36
0.088507991	-1.35
0.090122672	-1.34
0.091759136	-1.33
0.093417509	-1.32
0.095097918	-1.31
0.096800485	-1.30
0.098525329	-1.29
0.100272568	-1.28
0.102042315	-1.27

0.103834681	-1.26
0.105649774	-1.25
0.107487697	-1.24
0.109348552	-1.23
0.111232437	-1.22
0.113139446	-1.21
0.11506967	-1.20
0.117023196	-1.19
0.119000107	-1.18
0.121000484	-1.17
0.123024403	-1.16
0.125071936	-1.15
0.127143151	-1.14
0.129238112	-1.13
0.131356881	-1.12
0.133499513	-1.11
0.135666061	-1.10
0.137856572	-1.09
0.14007109	-1.08
0.142309654	-1.07
0.1445723	-1.06
0.146859056	-1.05
0.14916995	-1.04
0.151505003	-1.03
0.15386423	-1.02
0.156247645	-1.01
0.158655254	-1.00
0.16108706	-0.99
0.163543059	-0.98
0.166023246	-0.97
0.168527607	-0.96
0.171056126	-0.95
0.17360878	-0.94
0.176185542	-0.93
0.17878638	-0.92
0.181411255	-0.91
0.184060125	-0.90
0.186732943	-0.89
0.189429655	-0.88
0.192150202	-0.87
0.194894521	-0.86
0.197662543	-0.85
0.200454193	-0.84

0.203269392	-0.83
0.206108054	-0.82
0.208970088	-0.81
0.211855399	-0.80
0.214763884	-0.79
0.217695438	-0.78
0.220649946	-0.77
0.223627292	-0.76
0.226627352	-0.75
0.229649997	-0.74
0.232695092	-0.73
0.235762498	-0.72
0.238852068	-0.71
0.241963652	-0.70
0.245097094	-0.69
0.24825223	-0.68
0.251428895	-0.67
0.254626915	-0.66
0.257846111	-0.65
0.2610863	-0.64
0.264347292	-0.63
0.267628893	-0.62
0.270930904	-0.61
0.274253118	-0.60
0.277595325	-0.59
0.280957309	-0.58
0.284338849	-0.57
0.287739719	-0.56
0.291159687	-0.55
0.294598516	-0.54
0.298055965	-0.53
0.301531788	-0.52
0.305025731	-0.51
0.308537539	-0.50
0.312066949	-0.49
0.315613697	-0.48
0.319177509	-0.47
0.32275811	-0.46
0.32635522	-0.45
0.329968554	-0.44
0.333597821	-0.43
0.337242727	-0.42
0.340902974	-0.41

0.344578258	-0.40
0.348268273	-0.39
0.351972708	-0.38
0.355691245	-0.37
0.359423567	-0.36
0.363169349	-0.35
0.366928264	-0.34
0.370699981	-0.33
0.374484165	-0.32
0.378280478	-0.31
0.382088578	-0.30
0.385908119	-0.29
0.389738752	-0.28
0.393580127	-0.27
0.397431887	-0.26
0.401293674	-0.25
0.405165128	-0.24
0.409045885	-0.23
0.412935577	-0.22
0.416833837	-0.21
0.420740291	-0.20
0.424654565	-0.19
0.428576284	-0.18
0.432505068	-0.17
0.436440537	-0.16
0.440382308	-0.15
0.444329995	-0.14
0.448283213	-0.13
0.452241574	-0.12
0.456204687	-0.11
0.460172163	-0.10
0.464143607	-0.09
0.468118628	-0.08
0.47209683	-0.07
0.476077817	-0.06
0.480061194	-0.05
0.484046563	-0.04
0.488033527	-0.03
0.492021686	-0.02
0.496010644	-0.01
0.5	0.00
0.503989356	0.01
0.507978314	0.02

0.511966473	0.03
0.515953437	0.04
0.519938806	0.05
0.523922183	0.06
0.52790317	0.07
0.531881372	0.08
0.535856393	0.09
0.539827837	0.10
0.543795313	0.11
0.547758426	0.12
0.551716787	0.13
0.555670005	0.14
0.559617692	0.15
0.563559463	0.16
0.567494932	0.17
0.571423716	0.18
0.575345435	0.19
0.579259709	0.20
0.583166163	0.21
0.587064423	0.22
0.590954115	0.23
0.594834872	0.24
0.598706326	0.25
0.602568113	0.26
0.606419873	0.27
0.610261248	0.28
0.614091881	0.29
0.617911422	0.30
0.621719522	0.31
0.625515835	0.32
0.629300019	0.33
0.633071736	0.34
0.636830651	0.35
0.640576433	0.36
0.644308755	0.37
0.648027292	0.38
0.651731727	0.39
0.655421742	0.40
0.659097026	0.41
0.662757273	0.42
0.666402179	0.43
0.670031446	0.44
0.67364478	0.45

0.67724189	0.46
0.680822491	0.47
0.684386303	0.48
0.687933051	0.49
0.691462461	0.50
0.694974269	0.51
0.698468212	0.52
0.701944035	0.53
0.705401484	0.54
0.708840313	0.55
0.712260281	0.56
0.715661151	0.57
0.719042691	0.58
0.722404675	0.59
0.725746882	0.60
0.729069096	0.61
0.732371107	0.62
0.735652708	0.63
0.7389137	0.64
0.742153889	0.65
0.745373085	0.66
0.748571105	0.67
0.75174777	0.68
0.754902906	0.69
0.758036348	0.70
0.761147932	0.71
0.764237502	0.72
0.767304908	0.73
0.770350003	0.74
0.773372648	0.75
0.776372708	0.76
0.779350054	0.77
0.782304562	0.78
0.785236116	0.79
0.788144601	0.80
0.791029912	0.81
0.793891946	0.82
0.796730608	0.83
0.799545807	0.84
0.802337457	0.85
0.805105479	0.86
0.807849798	0.87
0.810570345	0.88

0.813267057	0.89
0.815939875	0.90
0.818588745	0.91
0.82121362	0.92
0.823814458	0.93
0.82639122	0.94
0.828943874	0.95
0.831472393	0.96
0.833976754	0.97
0.836456941	0.98
0.83891294	0.99
0.841344746	1.00
0.843752355	1.01
0.84613577	1.02
0.848494997	1.03
0.85083005	1.04
0.853140944	1.05
0.8554277	1.06
0.857690346	1.07
0.85992891	1.08
0.862143428	1.09
0.864333939	1.10
0.866500487	1.11
0.868643119	1.12
0.870761888	1.13
0.872856849	1.14
0.874928064	1.15
0.876975597	1.16
0.878999516	1.17
0.880999893	1.18
0.882976804	1.19
0.88493033	1.20
0.886860554	1.21
0.888767563	1.22
0.890651448	1.23
0.892512303	1.24
0.894350226	1.25
0.896165319	1.26
0.897957685	1.27
0.899727432	1.28
0.901474671	1.29
0.903199515	1.30
0.904902082	1.31

0.906582491	1.32
0.908240864	1.33
0.909877328	1.34
0.911492009	1.35
0.913085038	1.36
0.914656549	1.37
0.916206678	1.38
0.917735561	1.39
0.919243341	1.40
0.920730159	1.41
0.922196159	1.42
0.92364149	1.43
0.9250663	1.44
0.92647074	1.45
0.927854963	1.46
0.929219123	1.47
0.930563377	1.48
0.931887882	1.49
0.933192799	1.50
0.934478288	1.51
0.935744512	1.52
0.936991636	1.53
0.938219823	1.54
0.939429242	1.55
0.940620059	1.56
0.941792444	1.57
0.942946567	1.58
0.944082597	1.59
0.945200708	1.60
0.946301072	1.61
0.947383862	1.62
0.948449252	1.63
0.949497417	1.64
0.950528532	1.65
0.951542774	1.66
0.952540318	1.67
0.953521342	1.68
0.954486023	1.69
0.955434537	1.70
0.956367063	1.71
0.957283779	1.72
0.958184862	1.73
0.959070491	1.74

0.959940843	1.75
0.960796097	1.76
0.96163643	1.77
0.96246202	1.78
0.963273044	1.79
0.964069681	1.80
0.964852106	1.81
0.965620498	1.82
0.966375031	1.83
0.967115881	1.84
0.967843225	1.85
0.968557237	1.86
0.969258091	1.87
0.969945961	1.88
0.97062102	1.89
0.97128344	1.90
0.971933393	1.91
0.97257105	1.92
0.973196581	1.93
0.973810155	1.94
0.97441194	1.95
0.975002105	1.96
0.975580815	1.97
0.976148236	1.98
0.976704532	1.99
0.977249868	2.00
0.977784406	2.01
0.978308306	2.02
0.97882173	2.03
0.979324837	2.04
0.979817785	2.05
0.98030073	2.06
0.980773828	2.07
0.981237234	2.08
0.9816911	2.09
0.982135579	2.10
0.982570822	2.11
0.982996977	2.12
0.983414193	2.13
0.983822617	2.14
0.984222393	2.15
0.984613665	2.16
0.984996577	2.17

0.985371269	2.18
0.985737882	2.19
0.986096552	2.20
0.986447419	2.21
0.986790616	2.22
0.987126279	2.23
0.987454539	2.24
0.987775527	2.25
0.988089375	2.26
0.988396208	2.27
0.988696156	2.28
0.988989342	2.29
0.98927589	2.30
0.989555923	2.31
0.989829561	2.32
0.990096924	2.33
0.99035813	2.34
0.990613294	2.35
0.990862532	2.36
0.991105957	2.37
0.991343681	2.38
0.991575814	2.39
0.991802464	2.40
0.99202374	2.41
0.992239746	2.42
0.992450589	2.43
0.992656369	2.44
0.992857189	2.45
0.993053149	2.46
0.993244347	2.47
0.993430881	2.48
0.993612845	2.49
0.993790335	2.50
0.993963442	2.51
0.994132258	2.52
0.994296874	2.53
0.994457377	2.54
0.994613854	2.55
0.994766392	2.56
0.994915074	2.57
0.995059984	2.58
0.995201203	2.59
0.995338812	2.60

0.995472889	2.61
0.995603512	2.62
0.995730757	2.63
0.995854699	2.64
0.995975411	2.65
0.996092967	2.66
0.996207438	2.67
0.996318892	2.68
0.996427399	2.69
0.996533026	2.70
0.99663584	2.71
0.996735904	2.72
0.996833284	2.73
0.996928041	2.74
0.997020237	2.75
0.997109932	2.76
0.997197185	2.77
0.997282055	2.78
0.997364598	2.79
0.99744487	2.80
0.997522925	2.81
0.997598818	2.82
0.9976726	2.83
0.997744323	2.84
0.997814039	2.85
0.997881795	2.86
0.997947641	2.87
0.998011624	2.88
0.998073791	2.89
0.998134187	2.90
0.998192856	2.91
0.998249843	2.92
0.99830519	2.93
0.998358939	2.94
0.99841113	2.95
0.998461805	2.96
0.998511001	2.97
0.998558758	2.98
0.998605113	2.99
0.998650102	3.00
0.998693762	3.01
0.998736127	3.02
0.998777231	3.03

0.998817109	3.04
0.998855793	3.05
0.998893315	3.06
0.998929706	3.07
0.998964997	3.08
0.998999218	3.09
0.999032397	3.10
0.999064563	3.11
0.999095745	3.12
0.999125968	3.13
0.999155261	3.14
0.999183648	3.15
0.999211154	3.16
0.999237805	3.17
0.999263625	3.18
0.999288636	3.19
0.999312862	3.20
0.999336325	3.21
0.999359047	3.22
0.999381049	3.23
0.999402352	3.24
0.999422975	3.25
0.999442939	3.26
0.999462263	3.27
0.999480965	3.28
0.999499063	3.29
0.999516576	3.30
0.99953352	3.31
0.999549913	3.32
0.99956577	3.33
0.999581108	3.34
0.999595942	3.35
0.999610288	3.36
0.999624159	3.37
0.999637571	3.38
0.999650537	3.39
0.999663071	3.40
0.999675186	3.41
0.999686894	3.42
0.999698209	3.43
0.999709143	3.44
0.999719707	3.45
0.999729912	3.46

0.999739771	3.47
0.999749293	3.48
0.99975849	3.49
0.999767371	3.50
0.999775947	3.51
0.999784227	3.52
0.99979222	3.53
0.999799936	3.54
0.999807384	3.55
0.999814573	3.56
0.999821509	3.57
0.999828203	3.58
0.999834661	3.59
0.999840891	3.60
0.999846901	3.61
0.999852698	3.62
0.999858289	3.63
0.999863681	3.64
0.99986888	3.65
0.999873892	3.66
0.999878725	3.67
0.999883383	3.68
0.999887873	3.69
0.9998922	3.70
0.99989637	3.71
0.999900389	3.72
0.99990426	3.73
0.99990799	3.74
0.999911583	3.75
0.999915043	3.76
0.999918376	3.77
0.999921586	3.78
0.999924676	3.79
0.999927652	3.80
0.999930517	3.81
0.999933274	3.82
0.999935928	3.83
0.999938483	3.84
0.999940941	3.85
0.999943306	3.86
0.999945582	3.87
0.999947772	3.88
0.999949878	3.89

0.999951904	3.90
0.999953852	3.91
0.999955726	3.92
0.999957527	3.93
0.999959259	3.94
0.999960924	3.95
0.999962525	3.96
0.999964064	3.97
0.999965542	3.98
0.999966963	3.99
0.999968329	4.00
0.999969641	4.01
0.999970901	4.02
0.999972112	4.03
0.999973274	4.04
0.999974391	4.05
0.999975464	4.06
0.999976493	4.07
0.999977482	4.08
0.999978431	4.09
0.999979342	4.10
0.999980217	4.11
0.999981056	4.12
0.999981862	4.13
0.999982635	4.14
0.999983376	4.15
0.999984088	4.16
0.99998477	4.17
0.999985425	4.18
0.999986052	4.19
0.999986654	4.20
0.999987231	4.21
0.999987785	4.22
0.999988315	4.23
0.999988824	4.24
0.999989311	4.25
0.999989779	4.26
0.999990226	4.27
0.999990655	4.28
0.999991066	4.29
0.99999146	4.30
0.999991837	4.31
0.999992199	4.32

0.999992545	4.33
0.999992876	4.34
0.999993193	4.35
0.999993497	4.36
0.999993788	4.37
0.999994066	4.38
0.999994332	4.39
0.999994587	4.40
0.999994831	4.41
0.999995065	4.42
0.999995288	4.43
0.999995502	4.44
0.999995706	4.45
0.999995902	4.46
0.999996089	4.47
0.999996268	4.48
0.999996439	4.49
0.999996602	4.50
0.999996759	4.51
0.999996908	4.52
0.999997051	4.53
0.999997187	4.54
0.999997318	4.55
0.999997442	4.56
0.999997561	4.57
0.999997675	4.58
0.999997784	4.59
0.999997888	4.60
0.999997987	4.61
0.999998081	4.62
0.999998172	4.63
0.999998258	4.64
0.99999834	4.65
0.999998419	4.66
0.999998494	4.67
0.999998566	4.68
0.999998634	4.69
0.999998699	4.70
0.999998761	4.71
0.999998821	4.72
0.999998877	4.73
0.999998931	4.74
0.999998983	4.75

0.999999032	4.76
0.999999079	4.77
0.999999124	4.78
0.999999166	4.79
0.999999207	4.80
0.999999245	4.81
0.999999282	4.82
0.999999317	4.83
0.999999351	4.84
0.999999383	4.85
0.999999413	4.86
0.999999442	4.87
0.99999947	4.88
0.999999496	4.89
0.999999521	4.90
0.999999545	4.91
0.999999567	4.92
0.999999589	4.93
0.999999609	4.94
0.999999629	4.95
0.999999648	4.96
0.999999665	4.97
0.999999682	4.98
0.999999698	4.99
0.999999713	5.00

T Table

Note: Original file is tagged screen reader navigable pdf

**t Table**

cum. prob	<i>t</i> <sub>.50</sub>	<i>t</i> <sub>.75</sub>	<i>t</i> <sub>.80</sub>	<i>t</i> <sub>.85</sub>	<i>t</i> <sub>.90</sub>	<i>t</i> <sub>.95</sub>	<i>t</i> <sub>.975</sub>	<i>t</i> <sub>.99</sub>	<i>t</i> <sub>.995</sub>	<i>t</i> <sub>.999</sub>	<i>t</i> <sub>.9995</sub>
one-tail	<b>0.50</b>	<b>0.25</b>	<b>0.20</b>	<b>0.15</b>	<b>0.10</b>	<b>0.05</b>	<b>0.025</b>	<b>0.01</b>	<b>0.005</b>	<b>0.001</b>	<b>0.0005</b>
two-tails	<b>1.00</b>	<b>0.50</b>	<b>0.40</b>	<b>0.30</b>	<b>0.20</b>	<b>0.10</b>	<b>0.05</b>	<b>0.02</b>	<b>0.01</b>	<b>0.002</b>	<b>0.001</b>
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
<b>Z</b>	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	<b>Confidence Level</b>										

F Table

Note: Original file is tagged screen reader navigable pdf

### A.3. Critical values of the F-distribution

		df (numerator)										
		1	2	3	4	5	6	7	8	9	10	
df (denominator)	p											
	1	.05	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
		.01	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85
	2	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
		.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
	3	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
		.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
	4	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
		.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
	5	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
		.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
	6	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
		.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
	7	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
		.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
	8	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
		.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
	9	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
		.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	
11	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	
	.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	
12	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	
13	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	
14	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	
15	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	
16	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	
	.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	
17	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	
18	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	

## REFERENCES

- Aldrich, K. M., Hellier, E. J., & Edworthy, J. (2009). What determines auditory similarity? The effect of stimulus group and methodology. *The Quarterly Journal of Experimental Psychology*, 62(1), 63-83.
- Andre, A. D., & Cutler, H. A. (1998). *Displaying uncertainty in advanced navigation systems*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Annett, J. (1966). Training for perceptual skills. *Ergonomics*, 9(6), 459-468.
- Annett, J., & Paterson, L. (1967). Training for auditory detection. *Acta psychologica*, 27, 420-426.
- The AP Statistics Exam. (2017). Retrieved from [http://apcentral.collegeboard.com/apc/members/exam/exam\\_information/8357.html](http://apcentral.collegeboard.com/apc/members/exam/exam_information/8357.html)
- Bakan, D. (1966). The test of significance in psychological research. *Psychological bulletin*, 66(6), 423.
- Ballas, J. A. (1994). *Effect of event variations and sound duration on identification of everyday sound*. Paper presented at the International Community on Auditory Display, Santa Fe, New Mexico.
- Ballou, G. (2013). *Handbook for sound engineers*: Taylor & Francis.
- Batanero, C., Burrill, G., & Reading, C. (2011). *Teaching statistics in school mathematics- challenges for teaching and teacher education: A joint ICMI/IASE study: the 18th ICMI study* (Vol. 14): Springer Science & Business Media.

- Batterman, J. M., Schuett, J. H., & Walker, B. N. (2013). *Auditory displays for accessible fantasy sports*. Paper presented at the Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility, Bellevue, Washington.
- Batterman, J. M., & Walker, B. N. (2012a). *Displaying error & uncertainty in auditory graphs*. Paper presented at the Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility, Boulder, Colorado, USA.
- Batterman, J. M., & Walker, B. N. (2012b). *Pilot Data for Auditory Uncertainty Preferences*. Georgia Tech.
- Batterman, J. M., & Walker, B. N. (2013). *Auditory Graphs Need Error Bars: Validating Error-to-Sound Mappings and Scalings*. Paper presented at the International Conference on Auditory Display, Lodz, Poland.
- Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological methods, 10*(4), 389.
- Bennett, K. B., Toms, M. L., & Woods, D. D. (1993). Emergent features and graphical elements: Designing more effective configural displays. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 35*(1), 71-97.
- Berlyne, D. E., & Borsa, D. M. (1968). Uncertainty and the orientation reaction. *Perception & Psychophysics, 3*(1), 77-79.
- Beving, B., & Eblen, R. E. (1973). "Same" and "Different" Concepts and Children's Performance on Speech Sound Discrimination. *Journal of Speech, Language, and Hearing Research, 16*(3), 513-517.

- Beyth-Marom, R., Fidler, F., & Cumming, G. (2008). Statistical cognition: Towards evidence-based practice in statistics and statistics education. *Statistics Education Research Journal*, 7(2), 20-39.
- Bianchi, A., Oakley, I., & Kwon, D. S. (2012). Counting clicks and beeps: Exploring numerosity based haptic and audio PIN entry. *Interacting with computers*, 24(5), 409-422.
- Bisantz, A. M., Cao, D., Jenkins, M., Pennathur, P. R., Farry, M., Roth, E., . . . Pfautz, J. (2011). Comparing uncertainty visualizations for a dynamic decision-making task. *Journal of Cognitive Engineering and Decision Making*, 5(3), 277-293.
- Blalock Jr, H. M. (1987). Some general goals in teaching statistics. *Teaching Sociology*, 164-172.
- Blanche, P., & Merino, B. J. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, 39(3), 313-338.
- Bonebright, T. L. (2005). *A suggested agenda for auditory graph research*. Paper presented at the International Conference on Auditory Display, Limerick, Ireland.
- Bonebright, T. L. (2012). *Were those coconuts or horse hoofs? visual context effects on identification and veracity of everyday sounds*. Paper presented at the International Conference on Auditory Display, Atlanta, GA.
- Bonebright, T. L., & Miner, N. E. (2005). Evaluation of auditory displays: Comments on Bonebright et al., ICAD 1998. *ACM Transactions on Applied Perception (TAP)*, 2(4), 517-520.
- Bonebright, T. L., Nees, M. A., Connerley, T. T., & McCain, G. R. (2001). *Testing the effectiveness of sonified graphs for education: A programmatic research project*. Paper presented at the International Conference on Auditory Display, Espoo, Finland.

- Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*: MIT press.
- Brewster, S., Wright, P. C., & Edwards, A. D. (1995). Parallel earcons: Reducing the length of audio messages. *International Journal of Human-Computer Studies*, 43(2), 153-175.
- Brewster, S. A., Wright, P. C., & Edwards, A. D. (1993). *An evaluation of earcons for use in auditory human-computer interfaces*. Paper presented at the Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems.
- Brown, L. M., & Brewster, S. A. (2003). *Drawing by ear: Interpreting sonified line graphs*. Paper presented at the International Conference on Auditory Display, Glasgow, UK.
- Brown, L. M., Brewster, S. A., Ramloll, S., Burton, R., & Riedel, B. (2003). *Design guidelines for audio presentation of graphs and tables*. Paper presented at the International Conference on Auditory Display.
- Burt, J. L., Bartolome, D. S., Burdette, D. W., & Comstock Jr, J. (1995). A psychophysiological evaluation of the perceived urgency of auditory warning signals. *Ergonomics*, 38(11), 2327-2340.
- Chang, D., Nesbitt, K. V., & Wilkins, K. (2007). *The Gestalt principle of continuation applies to both the haptic and visual grouping of elements*. Paper presented at the EuroHaptics Conference, 2007 and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. World Haptics 2007. Second Joint.
- Cheatham, P. G., & White, C. T. (1954). Temporal numerosity: III. Auditory perception of number. *Journal of Experimental Psychology*, 47(6), 425.
- Chew, Y. C. (2014). *Assessing the Use of Auditory Graphs for Middle School Mathematics*. (Ph.D.), Georgia Tech, Atlanta, GA.

- Chew, Y. C. D., & Walker, B. N. (2013). *What did you say?: visually impaired students using bonephones in math class*. Paper presented at the Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility.
- Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387), 531-554.
- Coekin, J. (1970). An oscilloscope polar coordinate display for multi-dimensional data. *Radio and Electronic Engineer*, 40(2), 97.
- Correll, M., & Gleicher, M. (2013). *Error bars considered harmful*. Paper presented at the IEEE Visualization Poster Proceedings. IEEE.
- Correll, M., & Gleicher, M. (2014). Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12), 2142-2151.
- Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological methods*, 2(2), 161.
- Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Frontiers in psychology*, 1, 26.
- Cuddy, L. L., & Lunney, C. A. (1995). Expectancies generated by melodic intervals: Perceptual judgments of melodic continuity. *Perception & Psychophysics*, 57(4), 451-462.
- Cumming, G. (2009). Inference by eye: reading the overlap of independent confidence intervals. *Statistics in medicine*, 28(2), 205-220.

- Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *The Journal of cell biology*, 177(1), 7-11.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61(4), 532-574.
- Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist*, 60(2), 170.
- Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: where will the next mean fall? *Psychological methods*, 11(3), 217.
- Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, 3(4), 299-311.
- Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E., & Perrier, L. (2006). Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *Jama*, 296(9), 1094-1102.
- Davison, B. K. (2013). *Universal graph literacy: understanding how blind and low vision students can satisfy the common core standards with accessible auditory graphs*. (Ph.D.), Georgia Tech Atlanta, GA.
- Davison, B. K., & Walker, B. N. (2007). *Sonification Sandbox reconstruction: Software standard for auditory graphs*. Paper presented at the International Community on Auditory Displays, Montreal, Canada.
- de Campo, A., Frauenberger, C., Vogt, K., Wallisch, A., & Dayé, C. (2006). *Sonification as an interdisciplinary working process*. Paper presented at the International Conference on Auditory Display, London, UK.

- Deutsch, D., & Feroe, J. (1981). The internal representation of pitch sequences in tonal music. *Psychological review*, 88(6), 503.
- Distribution of Eligible Students Based on the Federal Quota Census of January 6, 2014 (Fiscal Year 2015). (2016).
- Doughty, J., & Garner, W. (1948). Pitch characteristics of short tones. II. Pitch as a function of tonal duration. *Journal of Experimental Psychology*, 38(4), 478.
- Equalisers Explained. (2001). *Sound on Sound*. Retrieved from <http://www.soundonsound.com/sos/jul01/articles/equalisers1.asp>
- Estes, W. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review*, 4(3), 330-341.
- Fidler, F., & Cumming, G. (2005). Teaching confidence intervals: Problems and potential solutions. *Proceedings of the 55th International Statistics Institute Session*.
- Fidler, F., & Cumming, G. (2007). Lessons learned from statistical reform efforts in other disciplines. *Psychology in the Schools*, 44(5), 441-449.
- Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., . . . Schmitt, R. (2005). Toward improved statistical reporting in the journal of consulting and clinical psychology. *Journal of Consulting and Clinical Psychology*, 73(1), 136.
- Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors Can Lead Researchers to Confidence Intervals, but Can't Make Them Think Statistical Reform Lessons From Medicine. *Psychological Science*, 15(2), 119-126.
- Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2005). Still Much to Learn About Confidence Intervals Reply to Rouder and Morey (2005). *Psychological Science*, 16(6), 494-495.

- Field, A. (2009). *Discovering statistics using SPSS*: Sage publications.
- Finger, R., & Bisantz, A. M. (2002). Utilizing graphical formats to convey uncertainty in a decision-making task. *Theoretical issues in ergonomics science*, 3(1), 1-25.
- Fisher, P. (1994). *Animation and sound for the visualization of uncertain spatial information*: Wiley Chichester,, UK.
- Flowers, J. H. (2005). Thirteen years of reflection on auditory graphing: Promises, pitfalls, and potential new directions. *Faculty Publications, Department of Psychology, University of Nebraska*, 430.
- Flowers, J. H., Buhman, D. C., & Turnage, K. D. (1997). Cross-modal equivalence of visual and auditory scatterplots for exploring bivariate data samples. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(3), 341-351.
- Flowers, J. H., & Grafel, D. C. (2002). *Perception of sonified daily weather records*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Flowers, J. H., & Hauer, T. A. (1992). The ear's versus the eye's potential to assess characteristics of numeric data: Are we too visuocentric? *Behavior Research Methods, Instruments, & Computers*, 24(2), 258-264.
- Flowers, J. H., & Hauer, T. A. (1993). "Sound" alternatives to visual graphics for exploratory data analysis. *Behavior Research Methods, Instruments, & Computers*, 25(2), 242-249.
- Flowers, J. H., & Hauer, T. A. (1995). Musical versus visual graphs: Cross-modal equivalence in perception of time series data. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(3), 553-569.

- Flowers, J. H., Whitwer, L. E., Grafel, D. C., & Kotan, C. A. (2001). *Sonification of daily weather records: Issues of perception, attention and memory in design choices*. Paper presented at the International Conference on Auditory Display.
- Frauenberger, C., Stockman, T., & Bourguet, M. L. (2007). *Pattern Design in the Context Space A Methodological Framework for Auditory Display Design*. Paper presented at the International Conference on Auditory Display, Montreal, Canada.
- Garfield, J. (1993). Teaching statistics using small-group cooperative learning. *Journal of Statistics Education, 1*(1), 1-9.
- Garner, W. (1951). The accuracy of counting repeated short tones. *Journal of Experimental Psychology, 41*(4), 310.
- Georgia Department of Education Statistical Reasoning Standards. (2017). Retrieved from <https://www.georgiastandards.org/Georgia-Standards/Frameworks/Statistical-Reasoning-Standards.pdf>
- Goldstein, M., Öquist, G., & Björk, S. (2002). Evaluating Sonified Rapid Serial Visual Presentation: An immersive reading experience on a mobile device *Universal Access Theoretical Perspectives, Practice, and Experience* (pp. 508-523): Springer.
- Grond, F., & Hermann, T. (2014). Interactive Sonification for Data Exploration: How listening modes and display purposes define design guidelines. *Organised Sound, 19*(1), 41.
- Guillaume, A., Pellieux, L., Chastres, V., & Drake, C. (2003). Judging the urgency of nonvocal auditory warning signals: perceptual and cognitive processes. *Journal of Experimental Psychology: Applied, 9*(3), 196.

- Hansen, E. G., Liu, L., Rogat, A., Hakkinen, M. T., & Darrah, M. (2016). Designing Innovative Science Assessments That Are Accessible for Students Who Are Blind. *Journal of Blindness Innovation & Research*, 6(1).
- Hellier, E., & Edworthy, J. (1989). Quantifying the perceived urgency of auditory warnings. *Canadian Acoustics*, 17(4), 3-11.
- Hermann, T. (2008). *Taxonomy and definitions for sonification and auditory display*. Paper presented at the International Conference on Auditory Display, Paris, France.
- Hermann, T., & Hunt, A. (2004). *The importance of interaction in sonification*. Paper presented at the International Conference on Auditory Display, Sydney, Australia.
- Hinders, D. C. (2015). *AP Statistics 2016*: McGraw Hill Education.
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5), 1157-1164.
- Horton, N. J., & Hardin, J. S. (2015). Teaching the Next Generation of Statistics Students to “Think With Data”: Special Issue on Statistics and the Undergraduate Curriculum. *The American Statistician*, 69(4), 259-265.
- Jones, M. R. (1976). Time, our lost dimension: toward a new theory of perception, attention, and memory. *Psychological review*, 83(5), 323.
- Joslyn, C., & Rocha, L. (1998). Towards a formal taxonomy of hybrid uncertainty representations. *Information Sciences*, 110(3), 255-277.
- Kalinowski, P. (2010). *Identifying misconceptions about confidence intervals*. Paper presented at the Proceedings of the Eighth International Conference on Teaching Statistics.[CDROM]. IASE, Ljubljana, Slovenia, Refereed paper.

- Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology*, 13(4), 351-371.  
doi:10.1002/(sici)1099-0720(199908)13:4<351::aid-acp589>3.0.co;2-6
- Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, 133(2), 189.
- Kaper, H. G., Tipei, S., & Wiebel, E. (1999). Data sonification and sound visualization. *Computing in science & engineering*, 1(4), 48-58.
- Kidd, G. R., Watson, C. S., & Gygi, B. (2007). Individual differences in auditory abilities. *The Journal of the Acoustical Society of America*, 122(1), 418-435.
- Kildal, J., & Brewster, S. A. (2005). *Explore the matrix: Browsing numerical data tables using sound*. Paper presented at the International Community on Auditory Display, Limerick, Ireland.
- Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, 3(3), 185-225.
- Kramer, G. (1993). *Auditory display: Sonification, audification, and auditory interfaces*: Perseus Publishing.
- Kramer, G., Walker, B. N., Bonebright, T. L., Cook, P., Flowers, J. H., Miner, N., & Neuhoff, J. (2010). *Sonification report: Status of the field and research agenda*. Paper presented at the International Conference on Auditory Display.
- Kryter, K. D. (2013). *The effects of noise on man*: Elsevier.

- Kryter, K. D., Ward, W. D., Miller, J. D., & Eldredge, D. H. (1966). Hazardous exposure to intermittent and steady-state noise. *The Journal of the Acoustical Society of America*, 39(3), 451-464.
- Krzywinski, M., & Altman, N. (2013). Points of significance: error bars. *Nature methods*, 10(10), 921-922.
- Lee, J. D., & Kirlik, A. (2013). *The Oxford handbook of cognitive engineering*: Oxford University Press.
- Leek, J. T., & Peng, R. D. (2015). Statistics: P values are just the tip of the iceberg. *Nature*, 520(7549), 612.
- Leslie, K. C., Low, R., Jin, P., & Sweller, J. (2012). Redundancy and expertise reversal effects when using educational technology to learn primary school science. *Educational Technology Research and Development*, 60(1), 1-13. doi:10.1007/s11423-011-9199-0
- Lodha, S. K., Wilson, C. M., & Sheehan, R. E. (1996). *LISTEN: sounding uncertainty visualization*. Paper presented at the Proceedings of the 7th conference on Visualization'96.
- Logue, A. (1976). Individual differences in magnitude estimation of loudness. *Perception & Psychophysics*, 19(3), 279-280.
- Ludovico, L. A., & Presti, G. (2016). The sonification space: A reference system for sonification tasks. *International Journal of Human-Computer Studies*, 85, 72-77.
- Madhyastha, T., & Reed, D. (1995). Data sonification: Do you see what I hear? *Software, IEEE*, 12(2), 45-56.
- Mancini, R., & Carter, B. (2009). *Op Amps for everyone*: Elsevier.

- Mathematics Georgia Standards of Excellence. (2017). Retrieved from <https://www.georgiastandards.org/Georgia-Standards/Pages/Math-9-12.aspx>
- Mathematics Standards. (2017). Retrieved from <http://www.corestandards.org/Math/>
- Mauney, L. M. (2006). *Individual differences in cognitive, musical, and perceptual abilities*. Citeseer.
- McNeer, R. R., Bohórquez, J., Özdamar, Ö., Varon, A. J., & Barach, P. (2007). A new paradigm for the design of audible alarms that convey urgency information. *Journal of clinical monitoring and computing*, 21(6), 353-363.
- Metatla, O., Bryan-Kinns, N., & Stockman, T. (2012). Interactive hierarchy-based auditory displays for accessing and manipulating relational diagrams. *Journal on Multimodal User Interfaces*, 5(3-4), 111-122.
- Metatla, O., Bryan-Kinns, N., Stockman, T., & Martin, F. (2015). *Sonification of reference markers for auditory graphs: Effects on non-visual point estimation tasks (2167-9843)*. Retrieved from
- Moore, D. S. (1988). Should Mathematicians Teach Statistics? *College Mathematics Journal*, 19(1), 3-7.
- Mulekar, M. S. (2015). *Cracking the AP Statistics Exam (2017 ed.)*: The Princeton Review.
- Nees, M. A., & Walker, B. N. (2006). *Relative intensity of auditory context for auditory graph design*. Paper presented at the International Conference on Auditory Display, London, UK.
- Nees, M. A., & Walker, B. N. (2007). *Listener, task, and auditory graph: Toward a conceptual model of auditory graph comprehension*. Paper presented at the International Conference on Auditory Display, Montreal, Canada.

- Nees, M. A., & Walker, B. N. (2008). Data density and trend reversals in auditory graphs: Effects on point-estimation and trend-identification tasks. *ACM Transactions on Applied Perception (TAP)*, 5(3), 13.
- Nees, M. A., & Walker, B. N. (2009). Auditory interfaces and sonification *The universal access handbook* (pp. 507-521).
- Neuhoff, J. G. (2004). *Ecological psychoacoustics*: Elsevier Academic Press.
- Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767), 333-380.
- Oren, M., Harding, C., & Bonebright, T. L. (2008). Design and Usability Testing of an Audio Platform Game for Players with Visual Impairments. *Journal of Visual Impairment & Blindness*, 102(12), 761.
- Padilla, L. M., Hansen, G., Ruginski, I. T., Kramer, H. S., Thompson, W. B., & Creem-Regehr, S. H. (2015). The influence of different graphical displays on nonexpert decision making under uncertainty. *Journal of Experimental Psychology: Applied*, 21(1), 37.
- Pang, A., Wittenbrink, C. M., & Lodha, S. K. (1997). Approaches to uncertainty visualization. *The Visual Computer*, 13(8), 370-390.
- Payton, M. E., Greenstone, M. H., & Schenker, N. (2003). Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance? *Journal of Insect Science*, 3(1), 34.
- Peres, S. C., Best, V., Brock, D., & Stockman, A. (2008). Auditory interfaces. *HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces*, 147-195.

- Peres, S. C., & Lane, D. M. (2003). *Sonification of statistical graphs*. Paper presented at the International Conference on Auditory Display, Boston, MA.
- Peres, S. C., & Lane, D. M. (2005). *Auditory graphs: The effects of redundant dimensions and divided attention*. Paper presented at the International Conference on Auditory Display, Limerick, Ireland.
- Pinker, S. (1990). A theory of graph comprehension. *Artificial intelligence and the future of testing*, 73-126.
- Ricketts, C., & Berry, J. (1994). Teaching statistics through resampling. *Teaching Statistics*, 16(2), 41-44.
- Robinson, D. W., & Dadson, R. S. (1956). A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, 7(5), 166.
- Rossing, T. D., & Stumpf, F. B. (1982). The Science of Sound. *American Journal of Physics*, 50, 955-955.
- Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological bulletin*, 57(5), 416.
- Schmidt, W. H., McKnight, C. C., & Raizen, S. (2007). *A splintered vision: An investigation of US science and mathematics education* (Vol. 3): Springer Science & Business Media.
- Schneider, B., Wright, A. A., Edelheit, W., Hock, P., & Humphrey, C. (1972). Equal loudness contours derived from sensory magnitude judgments. *The Journal of the Acoustical Society of America*, 51(6B), 1951-1959.
- Schnupp, J., Nelken, I., & King, A. (2011). *Auditory neuroscience: Making sense of sound*: MIT Press.

- Schuett, J. H., & Walker, B. N. (2013). *Measuring comprehension in sonification tasks that have multiple data streams*. Paper presented at the Proceedings of the 8th Audio Mostly Conference.
- Schuett, J. H., Winton, R. J., Batterman, J. M., & Walker, B. N. (2014). *Auditory weather reports: demonstrating listener comprehension of five concurrent variables*. Paper presented at the Proceedings of the 9th Audio Mostly: A Conference on Interaction With Sound, Aalborg, Denmark.
- Scown, H., Bartlett, M., & McCarley, J. S. (2014). *Statistically Lay Decision Makers Ignore Error Bars in Two-Point Comparisons*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, 14(1), 47-69.
- Smith, D. R., & Walker, B. N. (2002). *Tick-marks, axes, and labels: The effects of adding context to auditory graphs*. Paper presented at the International Conference on Auditory Display, Kyoto, Japan.
- Smith, D. R., & Walker, B. N. (2005). Effects of auditory context cues and training on performance of a point estimation sonification task. *Applied Cognitive Psychology*, 19(8), 1065-1087.
- Statistics Course Description. (2010). *The AP Program*. Retrieved from <https://secure-media.collegeboard.org/ap-student/course/ap-statistics-2010-course-exam-description.pdf>
- Sternstein, M. (2013). *AP Statistics* (7th ed.): Barron's Educational Series.
- Stevens, S. S. (1975). *Psychophysics*: Transaction Publishers.

- Stockman, T., Hind, G., & Frauenberger, C. (2005). *Interactive sonification of spreadsheets*. Paper presented at the International Conference on Auditory Display, Limerick, Ireland.
- Stockman, T., Nickerson, L. V., & Hind, G. (2005). *Auditory graphs: A summary of current experience and towards a research agenda*. Paper presented at the International Conference on Auditory Display, Limerick, Ireland.
- Stokes, A., Wickens, C. D., & Kite, K. (1990). Display technology-human factors concepts. *NASA STI/Recon Technical Report A, 91, 27333*.
- Student. (1908). The probable error of a mean. *Biometrika*, 1-25.
- Suzuki, Y., & Takeshima, H. (2004). Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America*, 116(2), 918-933.
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review*, 10(3), 251-296.  
doi:10.1023/a:1022193728205
- Teghtsoonian, M., & Teghtsoonian, R. (1971). How repeatable are Stevens's power law exponents for individual subjects? *Perception & Psychophysics*, 10(3), 147-149.
- Teghtsoonian, M., & Teghtsoonian, R. (1983). Consistency of individual exponents in cross-modal matching. *Perception & Psychophysics*, 33(3), 203-214.
- Teghtsoonian, R. (1971). On the exponents in Stevens' law and the constant in Ekman's law.
- Teghtsoonian, R. (1973). Range effects in psychophysical scaling and a revision of Stevens' law. *The American journal of psychology*, 3-27.
- ten Hoopen, G., & Vos, J. (1979). Effect on numerosity judgment of grouping of tones by auditory channels. *Perception & Psychophysics*, 26(5), 374-380.

- Trickett, S. B., & Trafton, J. G. (2006). Toward a comprehensive model of graph comprehension: Making the case for spatial cognition *Diagrammatic representation and inference* (pp. 286-300): Springer.
- Tukey, J. W. (1977). Exploratory data analysis.
- Turnage, K. D., Bonebright, T. L., Buhman, D. C., & Flowers, J. H. (1996). The effects of task demands on the equivalence of visual and auditory representations of periodic numerical data. *Behavior Research Methods, Instruments, & Computers*, 28(2), 270-274.
- Turnbull, W. W. (1944). Pitch discrimination as a function of tonal duration. *Journal of Experimental Psychology*, 34(4), 302.
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, 76(2), 105.
- Väljamäe, A., Steffert, T., Holland, S., Marimon, X., Benitez, R., Mealla, S., . . . Jordà, S. (2013). A review of real-time EEG sonification research.
- Walker, B. N. (2002). Magnitude estimation of conceptual data dimensions for use in sonification. *Journal of Experimental Psychology: Applied*, 8(4), 211.
- Walker, B. N. (2007). Consistency of magnitude estimations with conceptual data dimensions used for sonification. *Applied Cognitive Psychology*, 21(5), 579-599.
- Walker, B. N. (2013). *Sonification and auditory displays in electronic devices*. Paper presented at the Proceedings of the Sound and Music Computing Conference (SMC2013), Stockholm.
- Walker, B. N., & Cothran, J. T. (2003). *Sonification Sandbox: A graphical toolkit for auditory graphs*. Paper presented at the International Conference on Auditory Display, Boston, MA.

- Walker, B. N., & Kramer, G. (2004). Ecological psychoacoustics and auditory displays: Hearing, grouping, and meaning making. *Ecological psychoacoustics*, 150-175.
- Walker, B. N., Kramer, G., & Lane, D. M. (2000). *Psychophysical scaling of sonification mappings*. Paper presented at the International Conference on Auditory Display, Atlanta, GA.
- Walker, B. N., & Lane, D. M. (2001). *Psychophysical scaling of sonification mappings: a comparison of visually impaired and sighted listeners*. Paper presented at the International Community on Auditory Display, Espoo, Finland.
- Walker, B. N., & Mauney, L. M. (2004). *Individual differences, cognitive abilities, and the interpretation of auditory graphs*. Paper presented at the International Community on Auditory Display, Sydney, Australia.
- Walker, B. N., & Mauney, L. M. (2010). Universal design of auditory graphs: A comparison of sonification mappings for visually impaired and sighted listeners. *ACM Transactions on Accessible Computing (TACCESS)*, 2(3), 12.
- Walker, B. N., & Nees, M. A. (2005a). *An agenda for research and development of multimodal graphs*. Paper presented at the International Conference on Auditory Display, Limerick, Ireland.
- Walker, B. N., & Nees, M. A. (2005b). *Brief training for performance of a point estimation sonification task*. Paper presented at the International Conference on Auditory Display, Limerick, Ireland.
- Walker, B. N., & Nees, M. A. (2005c). *Conceptual versus perceptual training for auditory graphs*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

- Walker, B. N., & Nees, M. A. (2011). Theory of Sonification. In T. Hermann, A. Hunt, & J. G. Neuhoff (Eds.), *The Sonification Handbook*. Berlin, Germany: Logos Publishing House.
- Watson, C. S., & Kidd, G. R. (1994). *Factors in the design of effective auditory displays*. Paper presented at the International Community on Auditory Display, Santa Fe, New Mexico.
- Wertheimer, M. (1923). Laws of organization in perceptual forms. *A source book of Gestalt Psychology*.
- Wickens, C. D. (1992). *Engineering psychology and human performance*: HarperCollins Publishers.
- Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, 3(2), 159-177.
- Wickens, C. D., & Carswell, C. M. (1995). The proximity compatibility principle: its psychological foundation and relevance to display design. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(3), 473-494.
- Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). *Engineering psychology & human performance*: Psychology Press.
- Wittenbrink, C. M., Pang, A. T., & Lodha, S. K. (1996). Glyphs for visualizing uncertainty in vector fields. *Visualization and Computer Graphics, IEEE Transactions on*, 2(3), 266-279.
- Woods, D. D., Wise, J. A., & Hanes, L. F. (1981). *An evaluation of nuclear power plant safety parameter display systems*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Woolston, C. (2015). Psychology journal bans P values. *Nature*, 519, 9.

Yarbrough, C. (1987). The relationship of behavioral self-assessment to the achievement of basic conducting skills. *Journal of Research in Music Education*, 35(3), 183-189.